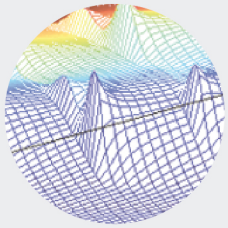
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM MATEMÁTICA APLICADA À ENGENHARIA E ÀS FINANÇAS






Clustering aplicado à Bolsa de Valores de Lisboa

CARLA SOFIA SOUSA JESUS
Outubro de 2015

POLITÉCNICO
DO PORTO

CARLA SOFIA SOUSA JESUS

Clustering aplicado à Bolsa de Valores de Lisboa

Dissertação para obtenção de grau de Mestre
Mestrado em Matemática aplicada à Engenharia e às Finanças

Outubro de 2015

Sob a orientação de Helena Cristina Mendes Brás

Agradecimentos

Gostaria de agradecer a todas as pessoas que contribuíram para que esta dissertação fosse concluída.

À minha orientadora, Professora Helena Brás, que acompanhou, sugeriu e corrigiu esta dissertação.

À diretora do Mestrado de Matemática Aplicada à Engenharia e às Finanças, Professora Stella Abreu, por todo o apoio durante todo o Mestrado.

Aos meus pais e irmã pelo apoio constante e por acreditarem sempre em mim.

Ao meu namorado, André, pelo incentivo, compreensão e encorajamento.

Índice

Resumo

O objetivo desta dissertação foi estudar um conjunto de empresas cotadas na bolsa de valores de Lisboa, para identificar aquelas que têm um comportamento semelhante ao longo do tempo. Para isso utilizamos algoritmos de *Clustering* tais como *K-Means, PAM*, Modelos hierárquicos, *Funny* e *C-Means* tanto com a distância euclidiana como com a distância de Manhattan. Para selecionar o melhor número de *clusters* identificado por cada um dos algoritmos testados, recorremos a alguns índices de avaliação/validação de *clusters* como o *Davies Bouldin* e *Calinski-Harabasz* entre outros.

Abstract

The aim of this thesis was to study a set of companies from Lisbon stock exchange to identify those that have a similar behavior over time. For this we use clustering algorithms such as K-Means, PAM, hierarchical models, Funny and C-Means with Euclidean distance and Manhattan distance. To select the best number of clusters identified by each of the tested algorithms, we resort to some clusters validation such as the Davies Bouldin and Calinski-Harabasz among others.

Índice de Tabelas

Índice de Figuras

1. Introdução

Data Mining é uma técnica para extrair informação útil de um conjunto de dados e é uma técnica muito utilizada pelos bancos e instituições financeiras com o objetivo de aumentar o desempenho dos seus negócios. Esta técnica é muito utilizada em mercado de ações para prever tendências futuras e descobrir padrões escondidos nestes mercados. As técnicas de *data mining* mais utilizadas no mercado de ações são árvores de decisão, redes neuronais, regras de associação, análise de fatores, *support vetor machines*, *Clustering* entre outros [1,2,3,4].

O objetivo de um investidor no mercado de ações é otimizar os retornos e para isso é necessário diminuir os riscos. Assim devem escolher ativos que têm comportamento diferente ao longo do tempo. Para isso pode-se usar os algoritmos de *Clustering*, primeiro para encontrar os ativos que têm comportamentos semelhantes ao longo do tempo que são definidos como *clusters* e por fim escolher elementos de *clusters* diferentes. A seleção de ativos do mercado de ações e a gestão de portfólios são problemas famosos em Finanças e existem diversos estudos nessa área, no entanto existem poucos estudos de *Clustering* aplicado a mercados de ações [5,3,6].

Cada empresa do mercado de ações é atribuída a um sector que melhor descreve a natureza do seu negócio. A evolução de um ativo depende de muitos fatores mas é esperado que as empresas dos mesmos sectores evoluam de forma semelhante ao longo do tempo. Assim alguns autores comparam os resultados dos algoritmos de *Clustering* com o *ground truth* que é o sector a que cada ativo pertence [7,8,9].

Dois dos diversos objetivos interessantes na aplicação de *Clustering* ao mercado de ações são otimização de portfólios [5,10] e verificar se empresas do mesmo setor têm o mesmo comportamento ao longo do tempo [7,9,11,12]. Para a otimização de um portfólio necessitamos de minimizar o risco do portfólio através da diversificação dos ativos [5,10].

Existem diversos algoritmos de *Clustering* aplicado a mercado de ações: desde o *k-means* [5,3,10], *SOM* [5], *FCM* [5], *PAM* ou *k-medoids* [7,10], *X-Means* [10], *DBSCAN* [3], modelos hierárquicos aglomerativos[11,8] e divisivos [7], TreeGNG [9], algoritmo

Gene Trajectorey Clustering [12] e um caso particular de modelo hierárquico com a distância *Hausdorff* [6].

Em relação aos índices para determinar qual é o melhor número de *clusters* dos resultados de *Clustering*, os autores de [5] usaram os índices *Silhoutte, Davies-Bouldin, Calinski-Lai, Dunn´s Index, Alternative Dunn's, Partition índex, Separation Index e Xie and Beni*, os autores de [7] apenas usaram o *Average silhouette width (ASW) e o Pearson version of Hubert's* (PH), em [3] usaram o *Davies Bouldin e o Dunn* e por fim os autores de [10] aplicaram *Silhouette e Davies Bouldin*.

Duas possíveis maneiras de avaliar os resultados dos algoritmos de *Clustering* (para além dos índices para determinar o número de *clusters*) são, ou através de uma técnica chamada *Intraclass inertia* [14], ou comparar os resultados com a *ground truth*. A técnica *Intraclass inertia* mede o quanto um *cluster* é compacto através da distância média entre cada elemento e o *centroid* do *cluster [14]*. Nos artigos [5,10] os autores utilizam a primeira opção e concluíram que *o k-means* obteve *clusters* mais compactos. Por outro lado, nos artigos [7,11,8,12] usaram a segunda maneira de avaliar os resultados. Em [7] os melhores resultados foram obtidos com o *PAM* em comparação com o *Agnes* e o *Diana* e em [11] o melhor método hierárquico foi o método *Ward* comparado com o *single linkage, o complete linkage e o average linkage*.

Enquanto os autores em [7,11,8] mantêm a opinião que empresas dos mesmos setores "movem-se juntas" ao longo do tempo, os autores de [12] referem que alguns sectores usualmente "movem-se juntos" como as empresas pertencentes à eletricidade e á área IT mas, por exemplo, os bancos normalmente não "movem-se juntos". Neste artigo [12] os autores estudam quais as empresas que se "movem juntas" em vários intervalos de tempo, ou seja, aplicam o método *Gene Trajectorey Clustering* a cada ano ao preço de fecho diário no período de 2000 a 2007.

Relativamente ao preço dos ativos que se pode usar para aplicar o *Clustering* existem autores que consideram o preço de abertura do ativo [8], autores que consideram o preço de fecho do ativo [6], outros consideram a média semanal do preço do ativo [11] e

ainda existem autores [12] que consideram o seguinte logaritmo: $log(p_i/p_{i-1})$ onde p_i é o preço de fecho do activo.

Séries Temporais são um conjunto de dados observados ao longo do tempo. A taxa de desemprego, o número de terramotos no mundo, as temperaturas máximas ou mínimas, as taxas de mortes ou nascimentos e o preço de ativos ao longo do ano são exemplos de séries temporais.

Clustering de séries temporais pode ser classificado em três abordagens: *Raw data based Approaches, Feature based Approaches* e *Model based Approaches*. A primeira abordagem é basicamente usar os dados originais sem qualquer alteração, a segunda abordagem usamos dados extraídos dos dados originais e na terceira abordagem são construídos modelos dos dados originais, como por exemplo os modelos AR, ARIMA ou Markov Chain [13].

O objetivo desta dissertação é utilizar vários algoritmos de *Clustering* para estudar quais são as empresas da bolsa de valores de Lisboa que têm comportamento semelhante ao longo do tempo. Também analisamos se essas empresas que têm comportamentos semelhantes ao longo do tempo pertencem aos mesmos setores ou não. O conjunto de dados estudado nesta dissertação é constituído por 38 empresas cotadas na bolsa de valores de Lisboa no período de 1 Janeiro de 2014 a 31 de Dezembro de 2014. Cada empresa é constituída por 255 valores que representam os preços de abertura das empresas.

O segundo capítulo desta dissertação é constituído por duas secções. A primeira secção é constituída pela definição de *Clustering* e pela descrição de vários algoritmos de *Clustering*. Os algoritmos definidos nesta secção são o *K-Means, PAM, Single Linkage, Complete Linkage, Average Linkage, Diana, Fuzzy C-Means (FCM)* e *Funny*. A segunda secção deste capítulo é constituída pela definição de validação de *Clusters* e pela descrição de vários índices para avaliar o número de *clusters* dos algoritmos da secção anterior.

O terceiro capítulo é constituído pela análise prática onde se aplica os algoritmos de *Clustering* e os índices definidos no segundo capítulo a um conjunto de dados de empresas cotadas na bolsa de valores de Lisboa.

Por fim, o quarto capítulo é constituído pelas conclusões desta dissertação assim como possíveis trabalhos futuros.

2. *Clustering* e *Validação de Clusters*

Neste capítulo definimos os conceitos de *Clustering* e Validação de *clusters*.

Na secção do *Clustering* descrevemos vários algoritmos de Partição, hierárquicos e *Fuzzy*. Os métodos de partição descritos são o *k-means* e *PAM*. Os métodos hierárquicos são o *single linkage, complete linkage, average linkage e o Diana*. Por fim, os métodos *fuzzy* descritos são o *fuzzy C-Means (FCM)* e o *Fanny*.

Na secção de validação de *clusters* descrevemos 13 *índices*: *Calinski-Harabasz, McClain Rao, C, Gamma, G Plus, Ray Turi, PBM, Davies-Bouldin, Point Biserial, Xie-Beni, Dunn, GDI e SD.*

2.1. *Clustering*

O processo de *Clustering* baseia-se em dividir um conjunto de dados em *clusters* de modo a que elementos do mesmo *cluster* sejam mais similares que elementos de *clusters* diferentes [15].

O processo de *Clustering* pode ser dividido em alguns passos fundamentais [16,17]:

- **Seleção de características**

  Na seleção de características analisamos o conjunto de dados e avaliamos se é necessário normalizar os dados, substituir valores em falta ou excluir dados não representativos.
  Uma normalização muito usada é a normalização no intervalo [0,1] e é definida através da seguinte fórmula:

$$\frac{x - x_{min}}{x_{max} - x_{min}}$$

  Os valores em falta podem ser tratados de várias maneiras. Uma dessas maneiras é simplesmente eliminar a variável que tem valores em falta mas só faz sentido recorrer a este método se existirem muitos valores em falta. Outra forma de resolver este problema é preencher os valores em falta com a média da variável.

- **Seleção de Algoritmos de *Clustering***

  Para além de decidir quais os algoritmos de *Clustering* é necessário escolher a medida que avalia a semelhança/dissemelhança e o número de *clusters* a considerar.

- **Validação dos Resultados**

  No processo de *clustering* é necessário avaliar os resultados com recurso à Validação de *clusters*, ou seja, utilizar os índices de validação para avaliar qual é o melhor número de clusters assim como o resultado dos algoritmos.

- **Interpretação dos Resultados**

  Por fim, é necessário saber um pouco da área em que se está a aplicar o *clustering* para poder interpretar os resultados.

Existem várias maneiras de classificar os algoritmos de *clustering*: podemos classifica-los com base no método adotado para definir os *clusters* ou se permite que os elementos pertençam a mais que um *cluster* [16].

Podemos classificar os algoritmos em métodos de Partição, métodos hierárquicos, métodos de *clustering* baseado em densidade ou *clustering* baseado em grelhas [16]. Nesta dissertação só vamos considerar os dois primeiros. A diferença entre métodos de partição e métodos hierárquicos é que o primeiro constrói partições e o segundo cria uma decomposição hierárquica.

Outra forma de classificar os métodos de *clustering* é em *fuzzy clustering* e *Crisp Clustering*. A diferença entre os dois métodos é que em *fuzzy clustering* cada elemento pode pertencer a mais que um *cluster* com um dado valor de associação e no *crisp clustering* cada elemento pertence a um e um só *cluster*. O algoritmo mais conhecido do *Fuzzy clustering* é o *Fuzzy C-Means*[16].

Clustering é uma área de constante desenvolvimento uma vez que é uma área que tem vários problemas sendo que o principal seja ser uma aprendizagem não supervisionada. Não existe um algoritmo de *clustering* que seja "bom" para qualquer tipo de conjunto de dados e assim é necessário aplicar vários algoritmos ao mesmo conjunto de dados. Outro problema é o número de *clusters* que por norma não sabemos à priori e necessitamos de decidir qual é o melhor e para isso usa-se índices de validação [18].

Por fim, da aplicação dos algoritmos de *Clustering* a qualquer conjunto de dados, obtém-se sempre uma partição dos mesmos mesmo que o conjunto dos dados não tenha qualquer estrutura [18].

2.1.1 Métodos de Partição

Os métodos de partição constroem uma partição de um conjunto de dados D com n elementos em k *clusters*. Dois algoritmos conhecidos destes métodos são o famoso *K-Means* e o *k-medoids*, também conhecido como *PAM (Partitioning Around Medoids)*. Ambos os algoritmos têm como objetivo minimizar a distância entre os elementos do *cluster* e o ponto designado como centro. A diferença entre os dois é como definem o centro, o *k-means* considera a média dos elementos (centroide) como o centro e o *PAM* considera o elemento que se encontra no centro do *cluster* (medoide) como o centro [17].

Os passos fundamentais do *k-means* estão representados a seguir[17]:

1. Escolhe aleatoriamente k elementos do conjunto de dados D como centros iniciais dos *clusters*
2. Atribui a cada elemento ao *cluster* com o centro mais próximo baseado na média dos elementos de cada *cluster*
3. Recalcula o centro de cada *cluster* atualizando a média dos elementos de cada *cluster*
4. Repete os passos 2 e 3 até não haver mais nenhuma mudança

Os passos fundamentais do algoritmo PAM estão representados a seguir[17]:

1. Escolhe aleatoriamente k elementos do conjunto de dados D como elementos representativos iniciais
2. Atribui cada elemento ao *cluster* com o elemento representativo mais perto
3. Escolhe aleatoriamente um elemento não considerado representativo
4. Calcula o custo (S) do elemento representativo e o elemento escolhido em 3
5. Se S<0 troca o elemento representativo com o escolhido em 3
6. Repete os passos anteriores até não haver mudanças

O *k-means* tem uma desvantagem em comparação com o *PAM* que é o facto de o centro dos *clusters* ser calculado através da média o que leva a que este algoritmo seja muito sensível a *outliers* e ruídos. No *PAM* o centro é calculado através do ponto mais centrado do *cluster* e por isso não é tao sensível a *outliers* como o *k-means*.
Por outro lado o PAM não é recomendado para conjunto de dados muito grandes e nesse cada usa-se o algoritmo *CLARA*. O algoritmo *CLARA* cria várias amostras do conjunto de dados e aplica o *PAM* a cada amostra e como *output* devolve o melhor *clustering* encontrado.

2.1.2 Métodos Hierárquicos

Os métodos hierárquicos constroem *clusters* através de uma estrutura hierárquica denominada de *dendograma*. Existem dois tipos de métodos hierárquicos, os aglomerativos e os divisivos. Os métodos aglomerativos começam o processo com cada elemento sendo um *cluster* e ao longo do processo esses clusters vão se juntando através de um dado critério até só existir um único cluster com todos os elementos. Os métodos divisivos é o contrário dos aglomerativos, começa o processo com todos os elementos pertencentes a um único *cluster* e ao longo do processo esses clusters vão se dividindo até só permanecer um cluster com todos os elementos [19,17,20].
Um algoritmo conhecido dos métodos hierárquicos divisivos é o *DIANA* e a descrição do algoritmo está descrita em baixo assim como o processo aglomerativo.

Algoritmo do Processo aglomerativo [20]

1. Representa cada elemento do conjunto de dados como um *cluster*
2. Calcula as distâncias entre *clusters* (Matriz de Proximidade)
3. Junta os *clusters* que têm a distância mínima
4. Atualiza a matriz de proximidade e repete os passos 2 e 3 até só existir um *cluster* com todos os elementos

Algoritmo DIANA [20]

Seja C_l o *cluster* que vai ser dividido nos clusters C_i e C_j.

1. Começa com C_i igual a C_l e C_j como um cluster vazio.
2. Para cada elemento que pertence a C_i:
    a. Na primeira iteração, calcula a sua distância média a todos os elementos.
    b. Para as iterações que faltam calcula a diferença entre a distância média até C_i e a distância média até C_j:
3.
    a. Para a primeira iteração, move os elementos com os valores maiores para o C_j;
    b. Para as restantes iterações, se o máximo valor da equação do 2.b. for maior que 0, move o elemento com diferença máxima para o C_j. Repete os passos 2.b e 3.b. e se não houver diferenças pára.

Os resultados do processo aglomerativo depende da distância usada no passo 2. Existem vários métodos para calcular a distância entre *cluster* mas nesta dissertação só vamos usar três: *Single Linkage, Complete Linkage e Average Linkage*.

A distância entre *clusters* determinada através do *single Linkage* é simplesmente a distância mínima entre dois pares de elementos que não pertencem ao mesmo *cluster* e por isso este método também é conhecido como método do vizinho mais próximo [19,21].

O método *Complete Linkage* também é conhecido por método do vizinho mais longe uma vez que a distância entre *clusters* é a distância máxima entre dois elementos que não pertencem ao mesmo *cluster* [19,21].

Por fim, o *método Average Linkage* como o nome indica, a distância entre dois *clusters* é a média das distâncias entre quaisquer pares de elementos que não pertencem ao mesmo *cluster* [19,21].

Num modelo hierárquico como obtemos um *dendrograma* como resultado final do processo obtemos partições múltiplas, ou seja, ao cortar o *dendrograma* em diferentes níveis obtemos várias partições o que pode ser visto como uma vantagem destes métodos[19].

Em relação às desvantagens destes métodos a principal tem a ver com o momento em que o método junta ou divide os elementos em *clusters*. Esta decisão não pode ser mudada e por isso não haverá troca de elementos dos *clusters* o que pode levar a uma qualidade de *clustering* menos boa se a escolha da junção/divisão não for a melhor. Assim, foram criados novos algoritmos para combater esta desvantagem. Estes algoritmos consistem em juntar modelos hierárquicos com outros tipos de *clustering*. Alguns desses algoritmos são *BIRCH, CURE, ROCK e CHAMELEON* [20].
Nesta dissertação só vamos usar os *linkage single, complete e average* dos algoritmos aglomerativos e o *DIANA* dos algoritmos divisivos.

2.1.3 Métodos *Fuzzy*

Fuzzy Clustering é um método que permite que cada elemento pertença a mais que um *cluster* com um nível/grau de associação. Geralmente, este tipo de método é mais natural que os métodos em que um elemento é obrigado a pertencer a apenas um *cluster*. Por exemplo, um elemento que está no centro do *cluster* tem um nível de associação diferente e maior que um elemento que está na fronteira com outro cluster. Assim também temos uma visão alargada de como os elementos estão distribuídos pelos *clusters.*

O algoritmo mais usado dos métodos *Fuzzy* é o *Fuzzy C-Means (FCM)* que foi introduzido por *Bezdek* em 1981 [22] e é considerado uma generalização do *ISODATA* [23].

FCM começa por calcular os centros de cada *clusters* e depois calcula o grau de associação de cada elemento para cada elemento nos clusters [24].

O objetivo do *FCM* é minimizar a seguindo função objetivo [20]:

$$J(U,M) = \sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^m D_{ij}^2$$

Onde

- c são os *fuzzy clusters* para um conjunto de dados $x_j, j = 1, \dots, N$
- $U = [u_{ij}]_{c\times N}$ é a matriz da partição *fuzzy*
- $u_{ij} \in [0,1]$ é o coeficiente de associação do elemento j no *cluster* i que satisfaz as seguintes restrições:

$$\sum_{i=1}^{c} u_{ij} = 1\,, \forall j$$

$$0 < \sum_{j=1}^{N} u_{ij} < N, \forall i$$

  A primeira restrição assegura que as associações totais de todos os *clusters* é 1 e a segunda restrição assegura que não exista *clusters* vazios.
- $M = [m_1, \dots, m_c]$ é a matriz do cluster protótipo (média ou centro)
- $m \in [1, \infty[$ é o parâmetro *fuzzifer* que é o parâmetro que controla a quantidade de sobreposição dos *clusters* [25], sendo assim um parâmetro muito importante. Normalmente usa-se m=2 [26]
- $D_{ij} = D(x_j, m_i)$ é a distância entre x_j e m_i e $D_{ij}^2 = \|x_j - m_i\|_2^2$

Os passos do algoritmo *FCM* são os seguintes [20]:

1. Seleciona os valores apropriados para m, c e um número pequeno positivo ε. Inicializa a matriz prototipa M aleatoriamente. Define a variável $t = 0$;
2. Atualiza da matriz de associação U da seguinte forma:

$$u_{ij}^{(t+1)} = \begin{cases} 1 \Big/ \left(\sum_{l=1}^{c} \left(\frac{D_{lj}}{D_{ij}} \right)^{2/(2-m)} \right), se\ I_j = \emptyset \\ 1/|I_j|, \quad Se\ I_j \neq \emptyset, i \in I_j \\ 0, I_j \neq \emptyset, i \notin I_j \end{cases}, para\ i = 1, \dots, c\ e\ j = 1, \dots, N$$

   Onde $I_j = \{i | i \in [1, c], x_j = m_i\}$

3. Atualiza a matriz protótipo M da seguinte forma:

$$m_i^{(t+1)} = \left(\sum_{j=1}^{N} \left(u_{ij}^{(t+1)} \right)^m x_j \right) \Big/ \sum_{j=1}^{N} \left(u_{ij}^{(t+1)} \right)^m \ para\ i = 1, \dots, c$$

4. Repete os passos 2 e 3 até $\left\| M^{(t+1)} - M^{(t)} \right\| < \varepsilon$

Como os métodos de partição também os métodos *fuzzy* têm a desvantagem, em comparação com os hierárquicos, que é necessário identificar o número de clusters e a partição inicial apropriada. Também os métodos *fuzzy* são influenciados pelo ruído e *outliers* [24,20,27].

Por outro lado, em comparação com o *k-means*, em [25] o autor refere que com o *fuzzy clustering* pode-se reduzir o número de mínimos locais. Mas a principal vantagem deste método é a de os elementos não pertencerem a só um *cluster* e assim representam melhor a realidade.

Outro algoritmo *fuzzy* que vamos aplicar na parte prática desta dissertação é o *FANNY*, este algoritmo é menos usado mas é um dos que está disponível no software R.
Kaufman e Rousseeuw em 1990 introduziram este algoritmo no livro *Finding Groups in Data* [28].

Como o algoritmo *FCM*, o objetivo do algoritmo *FUNNY* é minimizar uma função objetivo com base no nível de associação dos elementos nos *clusters* [29].

A função objetivo do algoritmo *FANNY* é [29]:

$$\sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 \, u_{jv}^2 d(i,j)}{2 \sum_{j=1}^{n} u_{jv}^2}$$

Onde

- u_{iv} é o nivel de associação do elemento i no *cluster* v e tem de satisfazer as seguintes condições:

$$u_{iv} \geq 0 \; para \; i = 1, \dots, n \; e \; v = 1, .., k$$

$$\sum_{v=1}^{k} u_{iv} = 1, para \; i = 1, \dots, n$$

- $d(i,j)$ é a distância entre os *clusters* i e j.

2.2 Validação de *Clusters*

Clustering é um método de aprendizagem não supervisionada e muitas das vezes não sabemos o número de clusters do conjunto de dados. Assim, necessitamos de uma maneira de descobrir e avaliar qual é o número de *clusters* apropriado para o conjunto de dados em estudo. Para isso usamos uma técnica chamada de validação de *clusters*.

Muitos trabalhos têm sido desenvolvidos com o objetivo de desenvolver métodos que permitam, além de identificar uma possível partição, também identificar o número de *clusters* mais adequado. Por exemplo, no artigo [56] os autores propuseram um algoritmo de *Clustering* não hierárquico baseado em coloração de grafos e um índice de validação baseado em *K-partite graphs (Clustering tendency índex).* O problema de coloração de grafos é um tema muito estudado em teoria de grafos e baseia-se em atribuir cores aos vértices do grafo de tal forma que dois vértices adjacentes tenham cores diferentes e o objetivo é minimizar o número de cores usadas. Ao conjunto de vértices da mesma cor chama-se classe de cores. Os autores de [56] começam por usar um algoritmo guloso para encontrar as classes de cores que são definidas como *clusters*

e depois otimizam esse resultado com o objetivo de melhorar o desempenho e identificar os *clusters* homogéneos.
O índice de validação proposto pelos autores deste artigo baseia-se na ideia que só existem *clusters* bem separados e compactos se for possível definir um *complete k-partite graph* do conjunto de dados depois de obter os resultados de *Clustering.*

Existem três tipos de validação de *clusters*: validação interna, externa e relativa. A diferença dos dois primeiros tipos é que o segundo usa informação à priori dos dados enquanto o primeiro usa apenas o conjunto de dados. Na validação externa uma informação necessária é o número de *clusters* do conjunto de dados o que nos problemas reais normalmente não é conhecido. Assim, a validação externa é apenas usada para avaliar e comparar os algoritmos de *clustering* enquanto a validação interna é usada não só para avaliar a qualidade dos algoritmos de *Clustering*, mas também para descobrir o número de *clusters* do conjunto de dados. A validação relativa avalia os resultados comparando esses resultados com outros cenários de *clusters*, ou seja usa vários algoritmos de *clustering* várias vezes com diferentes parâmetros de entrada [30,15,16]. Nesta dissertação só vamos utilizar a validação interna uma vez que não temos conhecimento de informação à priori dos dados.

Alguns índices baseiam-se na soma dos quadrados dentro dos *clusters* (SSW) e na soma dos quadrados entre *clusters* (SSB). O SSW avalia a dispersão dentro dos *clusters*, ou seja é o total das distâncias ao quadrado entre cada elemento e o centroide do *cluster*. O SSB avalia a dispersão dos *clusters*, ou seja calcula o total do quadrado das distâncias entre cada centroide do *cluster* ao centroide de todos os elementos. Alguns índices que pertencem a este grupo são: *Calinski e Harabasz, Hartigan, Ratkowsky e Lance* e *Ball e Hall* [31].

Neste capítulo definimos 13 índices que são os índices usados na parte prática desta dissertação e o critério para a escolha do melhor *cluster* para cada índice está definido na tabela 1.

De salientar que a escolha do melhor número de *clusters* não é o máximo/mínimo absoluto mas é escolhido através da maior/menor diferença entre números de clusters, chamado de "cotovelos".

Tabela 1 - Método e Critério para escolher o melhor número de clusters

Método	Critério
Calinski Harabasz	Max
Davies Bouldin	Min
C index	Min
Dunn	Max
Gamma	Max
G plus	Min
GDI	Max
McClain Rao	Min
PBM	Max
Point Biserial	Max
Ray Turi	Min
SD	Min
Xie Beni	Min

2.2.1 Índice *Calinski-Harabasz*

O Índice *Calinski-Harabasz* foi desenvolvido por *Calinski e Harabasz* em 1974 [32] e também é conhecido por Pseudo F Statistics [33].
Este índice é dado por [31]:

$$CH(k) = \frac{SSB/(k-1)}{SSW/(n-k)} = \frac{n-k}{k-1}\frac{SSB}{SSW}$$

Onde k é o número de *clusters* e n é o número do conjunto de dados.

O resultado desejado seria ter os *clusters* bem separados, ou seja um SSB elevado, e os elementos dentro dos *clusters* perto uns dos outros, ou seja um SSW pequeno. Assim como este índice é uma razão entre o SSB (numerador) e o SSW (denominador) um elevado valor deste índice é a melhor solução [34].

2.2.2 Índice *McClain Rao*

O índice *McClain e Rao* foi introduzido pelo *McClain e Rao* em 1975 [35] e consiste na razão entre dois termos. O numerador é a média da soma das distâncias dentro dos *clusters* dividido pelo número de distâncias entre pares que pertencem ao mesmo

cluster. O denominador é a média da soma das distâncias entre *clusters* dividido pelo total do número de distâncias entre pares que não pertencem ao mesmo *cluster* [36].

Este índice é então o quociente entre a média das distâncias dentro do *cluster* (distância *intra cluster*) e as distâncias entre *clusters* (distância *inter cluster*) [37,38]:

$$\frac{S_W/N_W}{S_B/N_B} = \frac{N_B}{N_W}\frac{S_W}{S_B}$$

2.2.3 Índice *C*

Hubert e Levin revisaram o índice C em 1976 [39] e também é conhecido como índice *Hubert e Levin* [31].

Este índice pode ser expresso na seguinte fórmula [31,40,36]:

$$\frac{d_w - \min(d_w)}{\max(d_w) - \min(d_w)}$$

Onde d_w é a soma das distâncias de todos os pares de elementos do mesmo *cluster*. Consideremos p como o número de pares de elementos no mesmo cluster. Assim, $\max(d_w)$ e $\min(d_w)$ são o máximo e o mínimo respetivamente das p menores e maiores distâncias de todas as distâncias entre pares de elementos [31,40].
Este índice pertence ao intervalo [0,1] e o valor mais pequeno deste índice indica o número ótimo de *clusters* [40].
Como este índice calcula e guarda todas as distâncias entre pares de elementos por vezes pode não ser possível aplicar este índice a conjunto de dados muito grandes [31].

2.2.4 Índice *Gamma*

Este índice foi introduzido por *Baker e Hubert* em 1975 [41] e é uma adapatação da estatística *Gamma de Godman e Kriskal's* para situações de *Clustering* [42,38].

A fórmula é a seguinte [40,36,42]:

$$Gamma = \frac{s^+ - s^-}{s^+ + s^-}$$

Onde

- s^+ é o número de comparações concordantes, ou seja, representa o número de vezes que a distância de dois pontos que não pertencem ao mesmo *cluster* é estritamente maior que que a distância entre dois pares de pontos que pertencem ao mesmo *cluster*
- s^- é o número de comparações discordantes, ou seja, representa o oposto de s^+, a distância entre dois pontos que não pertencem ao mesmo *cluster* é estritamente menor que a distância entre dois pontos que pertencem ao mesmo *cluster*.

2.2.5 Índice *G Plus*

Este índice foi inspecionado por *Rohlf* em 1974 [43] e usa apenas o conceito de comparações discordantes do índice *Gamma*.

Este índice é definido como os pares das distâncias dos dados que são discordantes normalizado pelo total das comparações das distâncias [36,37,38]:

$$Gplus = \frac{2s^-}{N_t(N_t - 1)}$$

Onde N_t é total dos pares de pontos distintos dos dados.

2.2.6 Indice *Ray Turi*

Ray e Turi propuseram em 1999 o índice *Ray Turi* que é baseado na razão entre a distância *intra cluster* e a distância *inter cluster* [44].

A distância *intra cluster* é a distância entre um elemento e o centro do seu *cluster* e depois é necessário fazer a média dessas distâncias [44]:

$$intra = \frac{1}{N}\sum_{i=1}^{k}\sum_{x \in C_i} \|x - z_i\|^2$$

Onde N é o número de elementos, k é o número de *clusters* e z_i é o centro do *cluster* C_i. A distância *inter cluster*, ou a distância entre os *clusters* é calculada através da distância entre os centros dos *clusters* e toma-se o mínimo desses valores [44].

$$inter = \min\left(\left\|z_i - z_j\right\|^2\right), i = 1,2, \dots, K-1, j = i+1, \dots, K$$

Assim, o índice *Ray Turi* pode ser definido da seguinte forma [44]:

$$RT = \frac{intra}{inter}$$

O objetivo é minimizar a distância *intra cluster*, que é o numerador do índice, e maximizar a distância *inter cluster*, que é o denominador do índice, assim podemos concluir que um valor baixo deste índice indica o número ideal de *clusters* [44].

2.2.7 Índice *PBM*

O índice *PBM* foi proposto por *Pakhira, Bandyopadhyay e Maulik* em 2003 e foi desenvolvido tanto para *crisp clustering* como para *fuzzy Clustering* [45].

Este índice é o produto de três fatores, definido da seguinte forma [37,45,46]:

$$PBM = \left(\frac{1}{K} \times \frac{E_1}{E_k} \times D_k\right)^2$$

Onde:

- K é o número de *clusters*
- E_1 é a soma de todas as distâncias dos elementos ao seu centro
- E_k é a soma das distâncias dentro dos *clusters* dos k *clusters*
- D_ké a maior distância entre dois centros de *clusters* e por isso indica a máxima separação entre pares de *clusters*

O primeiro fator deste índice, $\frac{1}{K}$, diminui com o aumento de k [45].

O segundo fator inclui a soma das distâncias *intra cluster* tomando os dados como um único *cluster* (E_1) e também para o k *cluster* (E_1). Neste factor o denominador decresce com o aumento de k e o numerador é fixo. Este numerador fixo é usado para evitar que este segundo fator seja muito pequeno. O objetivo será obter um valor elevado deste fator uma vez que este fator mede a compacidade do k *cluster* [45].

O terceiro fator, D_k mede a separação entre *clusters* e por isso queremos que este fator seja o maior possível [45].

Com o aumento de k, o primeiro fator diminui e os outros dois aumentam e isto deve se ao facto que queremos que o número de *clusters* seja pequeno mas também queremos aumentar a compacidade e a separação dos *clusters* [45].

2.2.8 Índice *Davies-Bouldin*

O índice *Davies Bouldin* foi proposto por *Davies e Bouldin* em 1978 [47] e é baseado na dispersão dentro dos *clusters* e na separação entre *clusters* [48].
A fórmula é a seguinte [40]:

$$DB = \frac{1}{n}\sum_{i=1,i\neq j}^{n} max\left(\frac{d_i + d_j}{d(c_i, c_j)}\right)$$

Onde n é o número de *clusters*, d_i e d_j são as médias da distância de todos os membros de cada *cluster* i ao respetivo centro c_i e c_j e $d(c_i, c_j)$ é a distância entre os centros dos *clusters*.
Um valor pequeno deste índice corresponde a *clusters* compactos e bem separados [40].

2.2.9 Índice *Point Biserial*

O coeficiente *Point Biserial* é usado em estatística como a medida de correlação entre uma variável A continua e uma variável binaria B. Ambas as variáveis tem o mesmo comprimento n [37].
O coefiente *Point Biserial* é definido da seguinte fórmula [37].

$$r_{pb}(A,B) = \frac{M_{A_1} - M_{A_0}}{s_n}\sqrt{\frac{n_{A_0} n_{A_1}}{n^2}}$$

Onde M_{A_0} e M_{A_1}são a média em A_0 eA_1 , n_{A_0} e n_{A_1} são os números de elementos de cada grupo e s_n é o desvio padrão de A.

Em termos de *clustering*, este coeficiente pode ser adaptado e a sua fórmula é a seguinte [37]:

$$PB = s_n \times r_{pb}(A,B) = \left(\frac{S_W}{N_W} - \frac{S_B}{N_B}\right)\frac{\sqrt{N_W N_B}}{N_T}$$

Onde A é as distâncias N_T entre os pares de pontos M_i e M_j e B é 1 se dois pontos pertencem ao mesmo *cluster* e 0 caso contrário [37].

$$A_{ij} = d(M_i, M_j)$$

$$B_{ij} = \begin{cases} 1\ se\ (i,j) \in I_W \\ 0\ caso\ contrario \end{cases}$$

M_{A_0} é constituída pelas média de todas as distâncias dentro do *cluster* e M_{A_1} é a média de todas as distâncias entre *clusters* [37].

A desvantagem deste índice é que pode falhar se os *clusters* tiverem tamanhos muito diferentes sendo assim desejável que os *clusters* tenham tamanhos semelhantes [49].

2.2.10 Índice *Xie-Beni*

O índice *Xie-Beni* foi introduzido por *Xie e Beni* em 1991 [50] e baseia-se na compacidade e separação dos *clusters* [16,20].
O índice pode ser definido da seguinte fórmula [50]:

$$XB = \frac{\pi}{s}$$

Onde π mede a compacidade dos *clusters*, ou seja, calcula a distância entre pares de *clusters* dentro do cluster e s mede a separação dos elementos em diferentes *clusters* e é definida como a distância mínima entre os centros dos *clusters* [16].

$$\pi = \sum_{j=1}^{c} \sum_{i=1}^{n} \frac{u_{ij}^2 \|x_i - v_j\|^2}{n}$$

$$s = {d_{min}}^2$$

d_{min} é a distância mínima entre os centros dos *clusters* dado por:

$$d_{min} = \min_{ij} \|v_i - v_j\|$$

Como pequenos valores de π indicam que os *clusters* são mais compactos e valores elevados de s indicam *clusters* bem separados então valores pequenos do índice XB significam *clusters* compactos e bem separados que é o objetivo de qualquer resultado de *clustering*.

2.2.11 Índice *Dunn*

O índice *Dunn* foi proposto por *Dunn* em 1974 [51] e baseia-se na distância menor do *intra-cluster* e na maior distância *inter-cluster* [37,30]:

$$D = \frac{d_{min}}{d_{max}}$$

d_{min} indica a menor distância entre dois elementos de diferentes *clusters* e d_{max} indica a maior distância entre dois elementos do mesmo *cluster*.
Duas desvantagens deste índice são que ele necessita de bastante tempo para calcular todas as distâncias necessárias [15,16] e é sensível ao ruído [15,16,52] mas existem generalizações deste índice que são mais robustas ao ruído (índices GDI).

2.2.12 Índice *GDI*

Bezdek em [53] introduziu generalizações do índice *Dunn* com o objetivo de melhorar a falha deste índice de ser sensível ao ruído. Estas generalizações também medem as distâncias entre *clusters* e dentro dos *clusters*. A fórmula geral é a seguinte [37]:

$$GDI = \frac{min_{k \neq k} \delta(C_k, C_{k'})}{max_k \Delta(C_k)}$$

Onde δ denota a medida da distância entre *clusters*, Δ denota a medida da distância dentro do *cluster* (que se chama o diâmetro do *cluster*) e $1 \leq k \leq K$ e $1 \leq k' \leq K$.

São definidos 6 diferentes definições para δ (denotadas por δ_1 até δ_6) e três definições para Δ (denotadas por Δ_1 até Δ_3). Assim temos 18 diferentes índices denotamos por C_{uv}: onde u é um número entre 1 e 6 e define a distância entre *clusters* e v é um número entre 1 e 3 e define a distância dentro dos grupos [37].
As definições das distâncias dentro dos *cluster* Δ são as seguintes [37]:

$$\Delta_1(C_k) = \max_{\substack{i,j \in I_k \\ i \neq k}} d(M_i, M_j)$$

$$\Delta_2(C_k) = \frac{1}{n_k(n_k - 1)} \sum_{\substack{i,j \in I_k \\ i \neq k}} d(M_i, M_j)$$

$$\Delta_3(C_k) = \frac{2}{n_k} \sum_{i \in I_k} d(M_i, G^{\{k\}})$$

onde d é a distância euclidiana e o factor 2 na definição de Δ_3 permite interpretar o valor como um diâmetro em vez de um raio.
As definições da distâncias entre *clusters* δ sãs as seguintes [37]:

$$\delta_1(C_k, C_{k'}) = \min_{\substack{i \in I_k \\ j \in I_{k'}}} d(M_i, M_j)$$

$$\delta_2(C_k, C_{k'}) = \max_{\substack{i \in I_k \\ j \in I_{k'}}} d(M_i, M_j)$$

$$\delta_3(C_k, C_{k'}) = \frac{1}{n_k n_{k'}} \sum_{\substack{i \in I_k \\ j \in I_{k'}}} d(M_i, M_j)$$

$$\delta_4(C_k, C_{k'}) = d(G^{\{k\}}, G^{\{k'\}})$$

$$\delta_5(C_k, C_{k'}) = \frac{1}{n_k + n_{k'}} \left(\sum_{i \in I_k} d\left(M_i, G^{\{k\}}\right) + \sum_{j \in I_{k'}} d\left(M_i, G^{\{k'\}}\right) \right)$$

$$\delta_6(C_k, C_{k'}) = max \left\{ \sup_{i \in I_k} \inf_{j \in I_{k'}} d\left(M_i, M_j\right), \sup_{j \in I_{k'}} \inf_{i \in I_k} d\left(M_i, M_j\right), \right\}$$

As quatro primeiras distâncias (δ_1 até δ_4) são chamadas de *single linkage, complete linkage, average linkage e centroid linkage*, respetivamente. δ_5 é a média ponderada (com pesos n_k e $n_{k'}$) das médias das distâncias entre pontos nos *clusters* C_k e $C_{k'}$ e seus respetivos centros e a medida δ_6 é a distância *Hausdorff* D_H [37].

2.2.13 Índice *SD*

O índice *SD* foi introduzido por *Halkidi, Batistakis e Vazirgiannis* em 2001 [54] e é baseado na média da dispersão dos *clusters* e na separação dos *clusters*. A fórmula é a seguinte [16,54]:

$$SD(n) = a.Scat(nc) + Dis(nc)$$

Onde

- Scat(nc) é a média da dispersão dos *clusters* e indica a compacidade média dos *clusters* (ou seja, a distância *inter cluster):*

$$Scat(nc) = \frac{1}{nc} \sum_{i=1}^{nc} \frac{\|\sigma(v_i)\|}{\sigma(X)}$$

- $Dis(nc)$ é a separação total dos *clusters* (ou seja, a distância *intra cluster*):

$$Dis(nc) = \frac{D_{max}}{D_{min}} \sum_{k=1}^{nc} \left(\sum_{z=1}^{nc} \|v_k - v_z\| \right)^{-1}$$

Onde

$D_{max} = \max(\|v_i - v_j\|) \ \forall i, j \in \{1,2,3, \dots, nc\}$ é a distância máxima entre os centros dos *clusters*.

$D_{min} = \min(\|v_i - v_j\|) \ \forall i, j \in \{1,2,3, \dots, nc\}$ é a distância mínima entre os centros dos *clusters*.

- a é o fator peso igual a $Dis(c_{max})$, onde c_{max} é o máximo de números de *clusters* possível.

Através das fórmulas deste índice podemos verificar que a distribuição dos centros influência a $Dis(nc)$ e consequentemente o índice *SD*. Outra medida que influência este índice é o número máximo de *clusters*. $Scat(nc)$ e $Dis(nc)$ são afectados pelo número de *clusters* através do fator de peso $a = Dis(c_{max})$ e consequentemente afeta o índice *SD* [54].

3. Análise de Clusters num Conjunto de Dados Financeiros

Este capítulo é constituído pela descrição do objetivo, do conjunto de dados utilizados, dos métodos/ algoritmos de *Clustering* usados, dos resultados obtidos e pelas conclusões deste capítulo.

3.1. Objetivo

O objetivo deste capítulo é estudar quais as empresas da bolsa de valores de Lisboa que têm um comportamento semelhante ao longo do tempo com recurso a alguns algoritmos de *Clustering* e a alguns índices para avaliar os resultados desses algoritmos, assim como avaliar o melhor número de *clusters*.

3.2. Conjunto de Dados

O conjunto de dados que estudamos é constituído por 38 empresas cotadas na bolsa de Lisboa de 1 de Janeiro de 2014 a 31 de Dezembro de 2014 [55]. Cada empresa tem 255 valores e foram escolhidos os preços de abertura das ações das empresas. Foram excluídas as empresas que tinham menos de 50% de valores e os valores em falta foram substituídos pela média de todos os valores conhecidos das empresas. O conjunto de dados foram normalizados entre 0 e 1.

O conjunto de dados é constituído por 38 séries temporais uma vez que são dados observados em diferentes instantes do tempo, neste caso diariamente.

As empresas estudadas nesta dissertação estão representadas na tabela 2 e na figura 1.

Tabela 2 - Empresas cotadas na bolsa de valores de Lisboa

ALTRI	EDP	LISGRAFICA	SEMAPA
BCP	EDP RENOVAVEIS	MARTIFER	SONAE
BPI	F.RAMA	MONTEPIO	SONAE CAPITAL
SANTANDER	FCP	MOTA ENGIL	SONAE IND.
BANIF	GALP	NOVABASE	SPORTING
BENFICA	GLINTT	P.TELECOM	SUMOL COMPAL
CIMPOR	IBERSOL	PORTUCEL	TEIXEIRA DUARTE
COFINA	IMPRESA	REN	TOYOTA CAETANO
CORTICEIRA AMORIM	INAPA	SAG GEST	
CTT	J.MARTINS	SDC INV.	

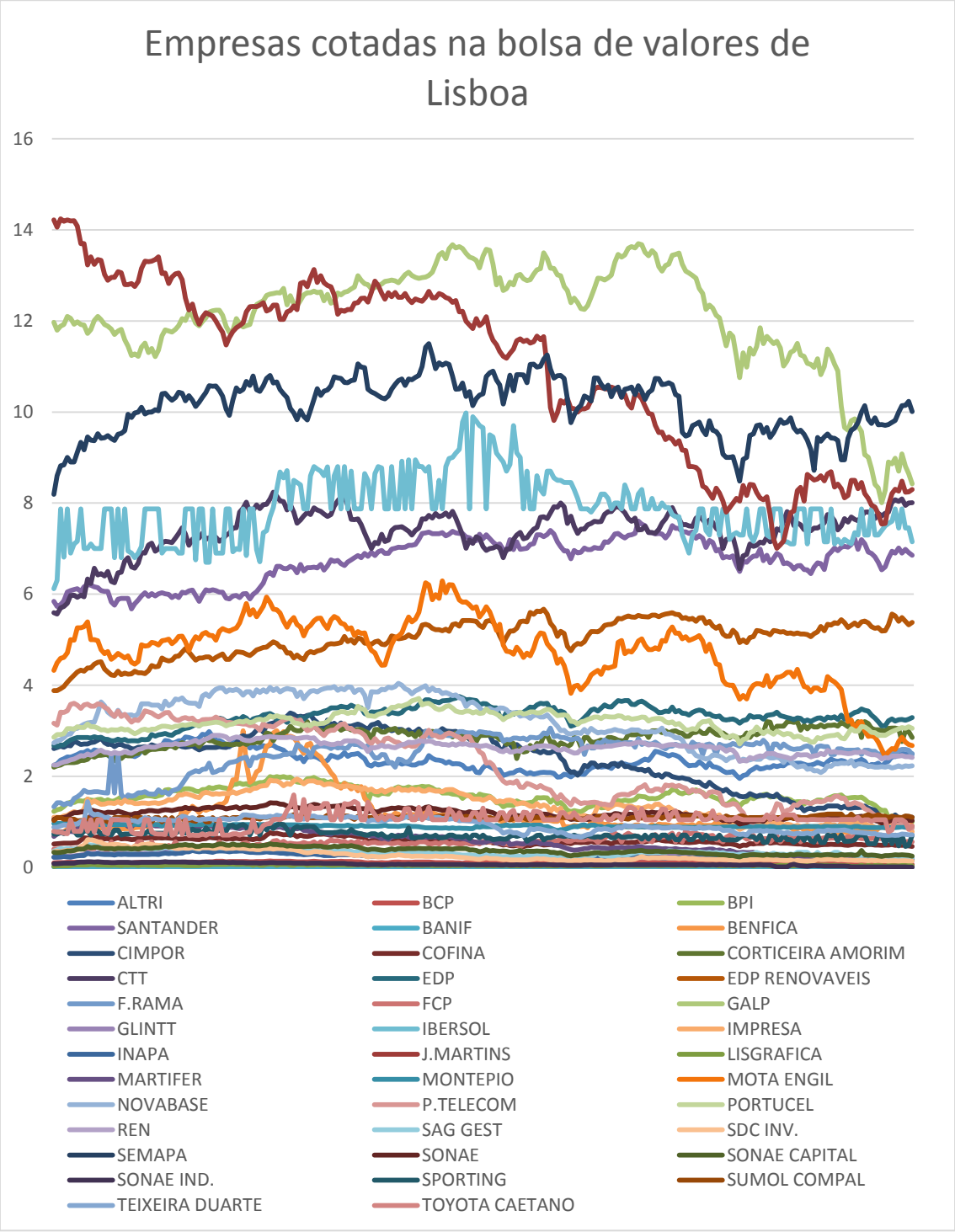


Figura 1 - Gráficos do conjunto de empresas estudadas nesta dissertação

3.3. Métodos usados

Os métodos/ algoritmos aplicados foram de três tipos: partição, hierárquicos e *fuzzy*. Os de partição que usamos foram o *K-Means* e o *PAM*. Os métodos hierárquicos usados nesta dissertação são três aglomerativos (*Complete, Single, Average*) e um divisivo (*Diana*). Por fim os métodos *fuzzy* utilizados foram o *C-Means* e o *Funny*. Em termos de distâncias usamos a euclidiana e a *Manhattan*. Para avaliar qual foi o melhor número de *clusters* para os dados em estudo usamos vários índices.

Nesta dissertação usou-se todos os índices do *Package ClusterCrit* do *software R* [37] mas como não obtivemos bons resultados com os índices *Banfeld Raftery*,*Ball Hall*, *Det Ratio, Ksq Det W*, *Log Det Ratio*,*Log SS Ratio, Ratkowsky Lance*, *Scott Symons*, *Tau*, *Trace W*, *Trace WiB e Wemmert Gançarski* estes índices foram excluídos deste estudo.
Como o melhor número de *clusters* para os vários *GDI* são "iguais" generalizamos estes índices para *GDI12*.

3.4. Resultados obtidos

Esta secção é constituída pelos resultados obtidos pelos algoritmos *K-Means, PAM*, Modelos Hierárquicos, *FCM* e *Funny*.
Para todos os algoritmos fizemos variar o número de *clusters* de 2 até 13 ($k = 2, \dots, 13$) e analisamos os vários resultados com diferentes índices.

3.4.1. K-Means

Depois de obter os resultados do algoritmo *K-Means* usamos vários índices para avaliar o melhor número de *clusters*.

Consideramos que o melhor número de *clusters* para o *k-means* com a distância euclidiana foi de 10 *clusters*, uma vez que foi considerada a melhor partição para 38% dos índices e com a distância de *Manhattan* foi de 9 *clusters* uma vez que foi considerada a melhor partição para 38% dos índices.

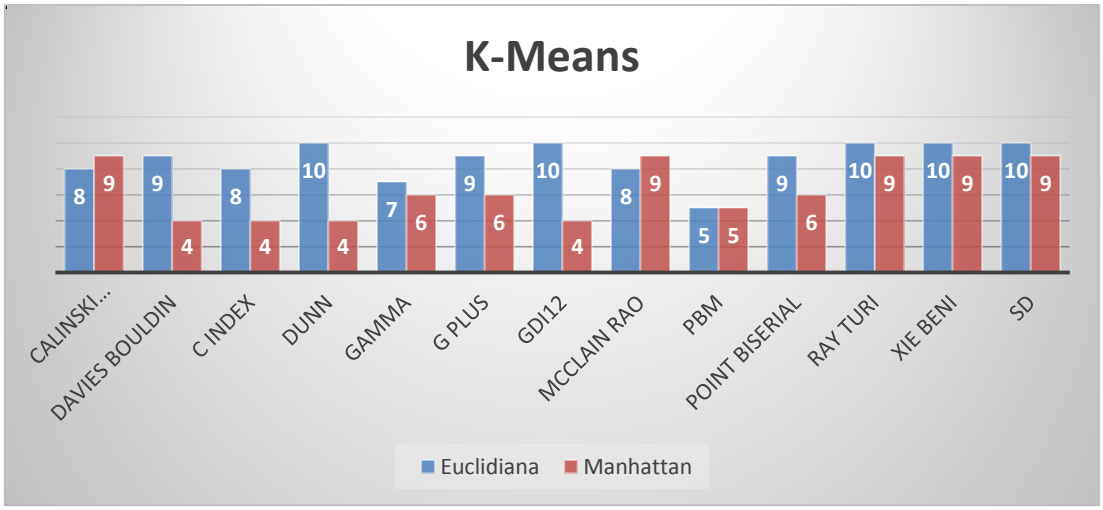


Figura 2 - Resultados dos índices pelo K-Means

Através dos índices *Dunn, GDI12, Ray Turi, Xie Beni e SD* obtemos o melhor número de *clusters* para o algoritmo *K-Means* com a distância euclidiana (10 *Clusters*).

Através dos índices *Calinski Harabasz, McClain Rao, Ray Turi, Xie Beni* e *SD* obtemos o melhor número de *clusters* para o *K-Means* com a distância de *Manhattan* (9 *clusters*).

Comparando os resultados do *k-means* com distância euclidiana e a distância de *Manhattan* os resultados com as duas distâncias foram diferentes sendo apenas com o *PBM* que obtivemos o mesmo resultado.

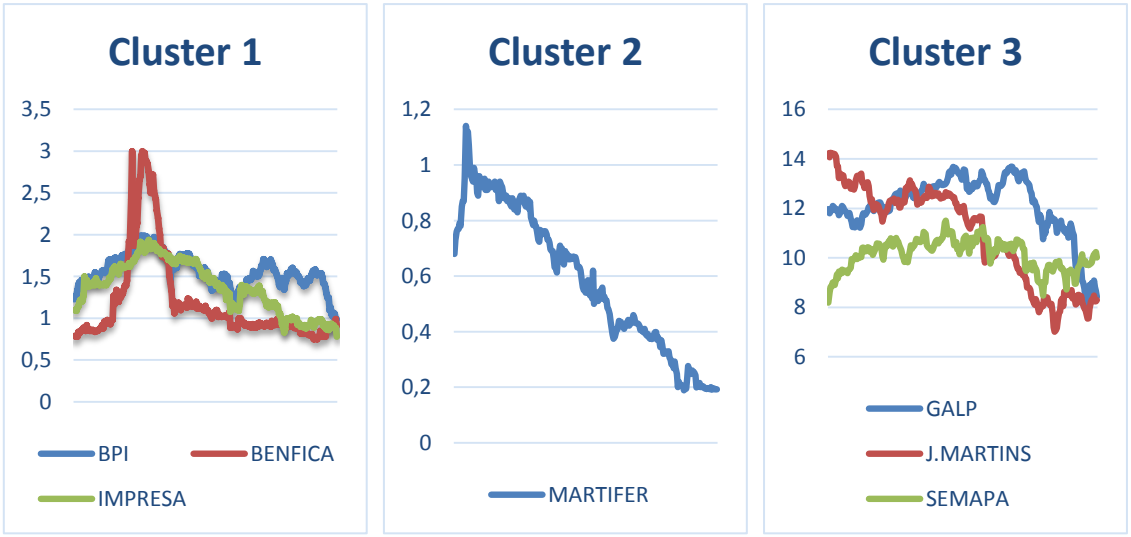


Figura 3 - Resultados dos clusters 1,2, e 3 do K-means/Euclidiana

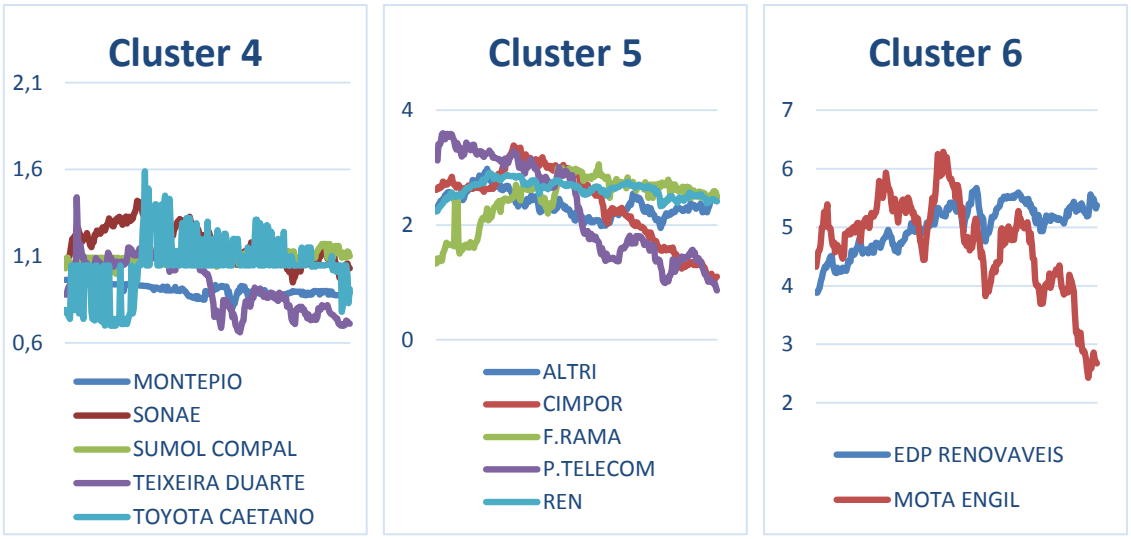


Figura 4 - Resultados dos clusters 4,5 e 6 do K-means/Euclidiana

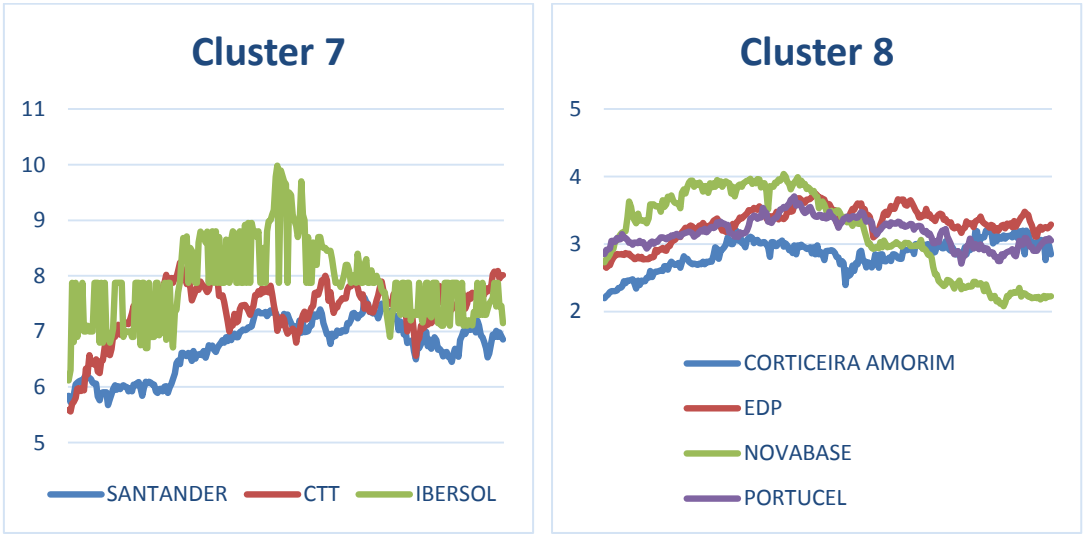


Figura 5 - Resultados dos clusters 7 e 8 do K-means/Euclidiana

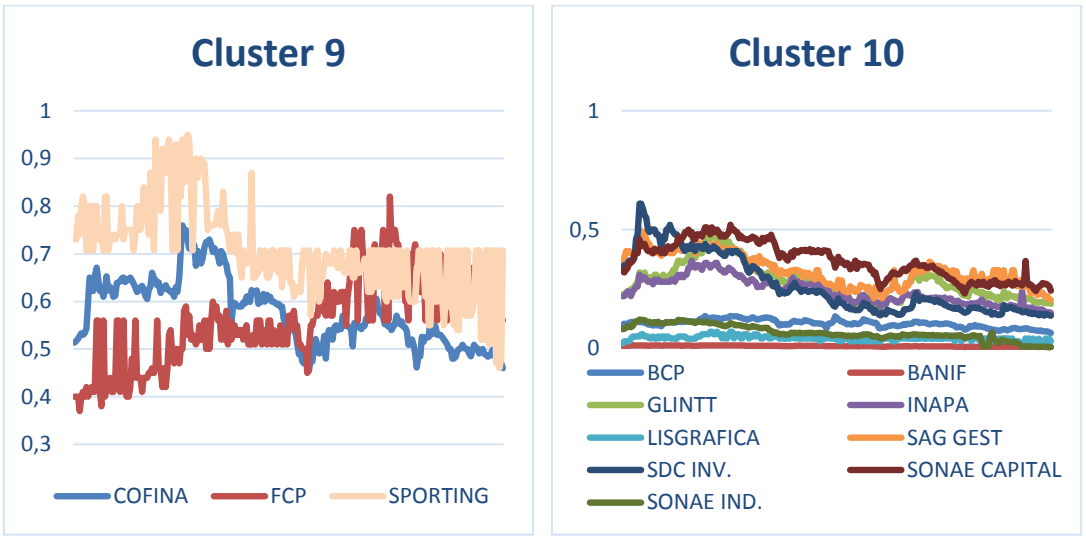


Figura 6 - Resultados dos clusters 9 e 10 do K-means/Euclidiana

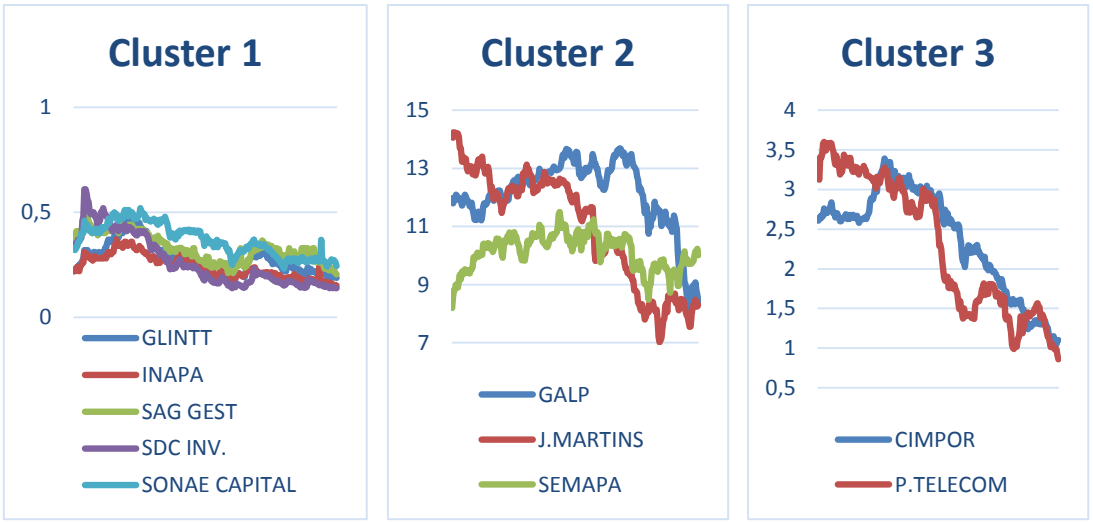


Figura 7 - Resultados dos clusters 1, 2 e 3 do K-means/Manhattan

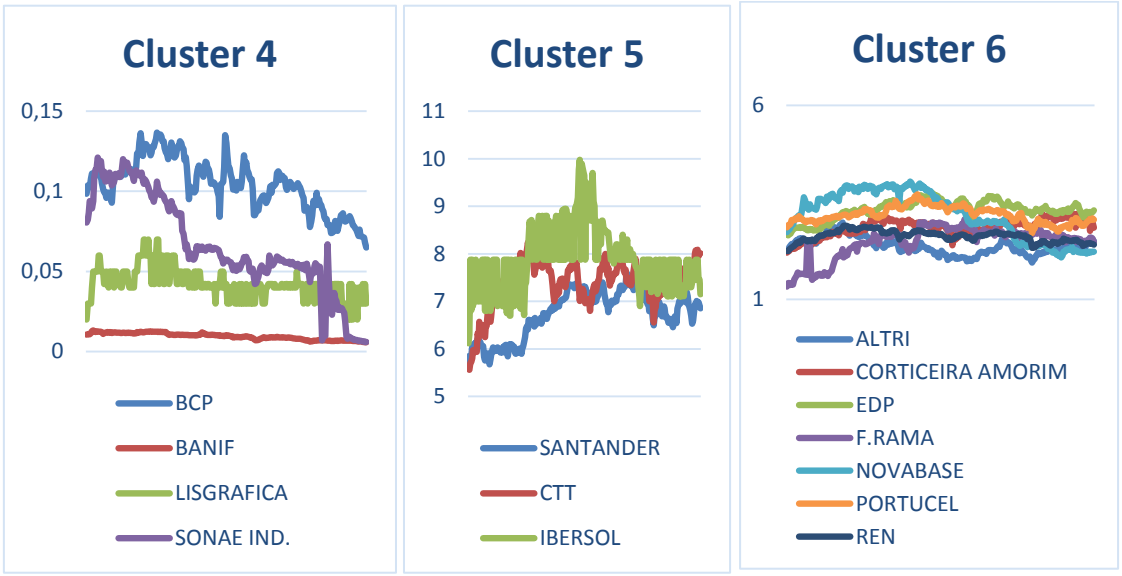


Figura 8 - Resultados dos clusters 4,5 e 6 do K-means/Manhattan

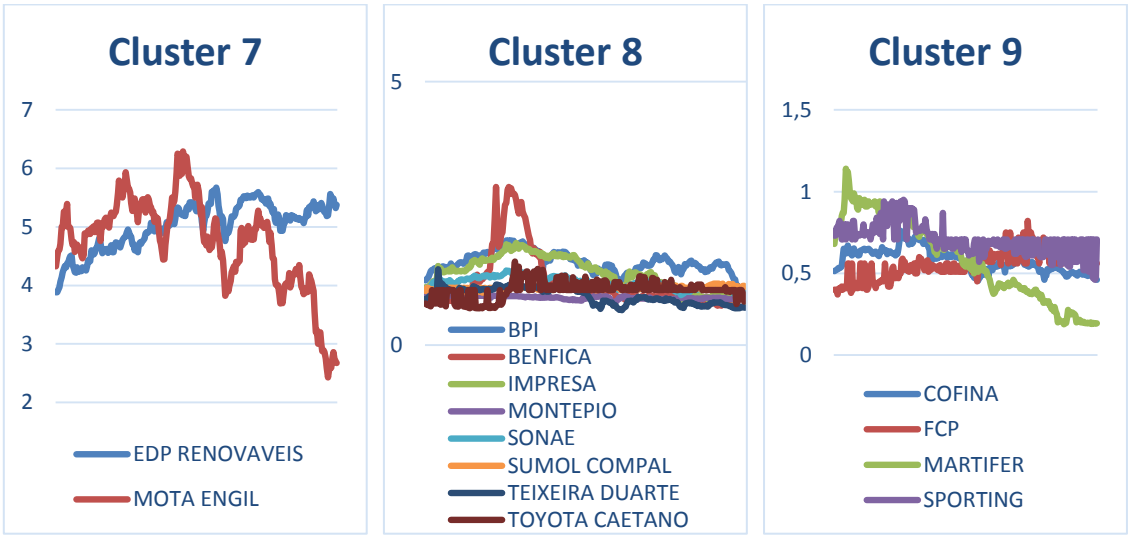


Figura 9 - Resultados dos clusters 7,8 e 9 do K-means/Manhattan

Analisando a constituição dos *clusters* das duas distâncias podemos concluir que existem *clusters* com as mesmas empresas. A Galp, J.Martins e Semapa pertencem ao mesmo *cluster* nas duas distâncias assim como o Santander, CTT e a Ibersol. Por fim, a Edp Renováveis e a Mota Engil são outras empresas que pertencem ao mesmo *cluster* tanto com a distância euclidiana como com a distância Manhattan.

3.4.2. PAM

Consideramos que o melhor número de *clusters* para o *PAM* com a distância euclidiana foi de 4 *clusters*, uma vez que foi considerada a melhor partição para 46% dos índices e com a distância de *Manhattan* foi de 4 *clusters* uma vez que foi considerada a melhor partição para 38% dos índices.

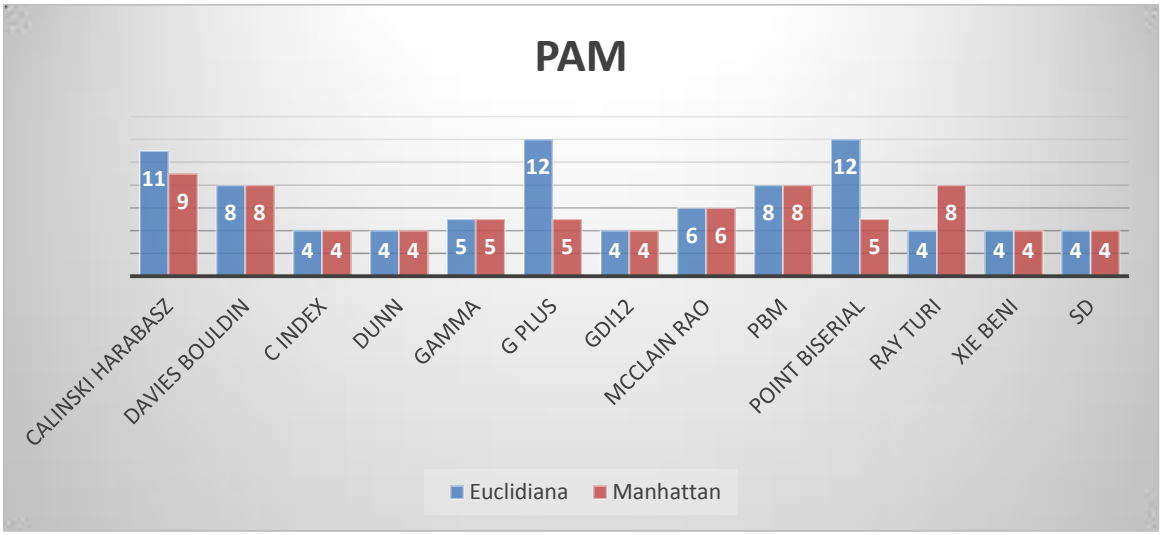


Figura 10 - Resultados dos índices pelo PAM

Através dos índices *C index*, *Dunn, GDI12, Ray Turi, Xie Beni* e *SD* obtemos o melhor número de *clusters* para o *PAM* com a distância de euclidiana (4 *clusters*).

Através dos índices *C índex*, *Dunn, GDI12, Xie Beni* e *SD* obtemos o melhor número de *clusters* para o *PAM* com a distância de *Manhattan*. (4 *clusters*).

Comparando os resultados do *PAM* com a distância euclidiana e a distância de *Manhattan*, com os índices *Calinski Harabasz, G plus, Point Biserial e Ray Turi* obtivemos resultados diferentes.

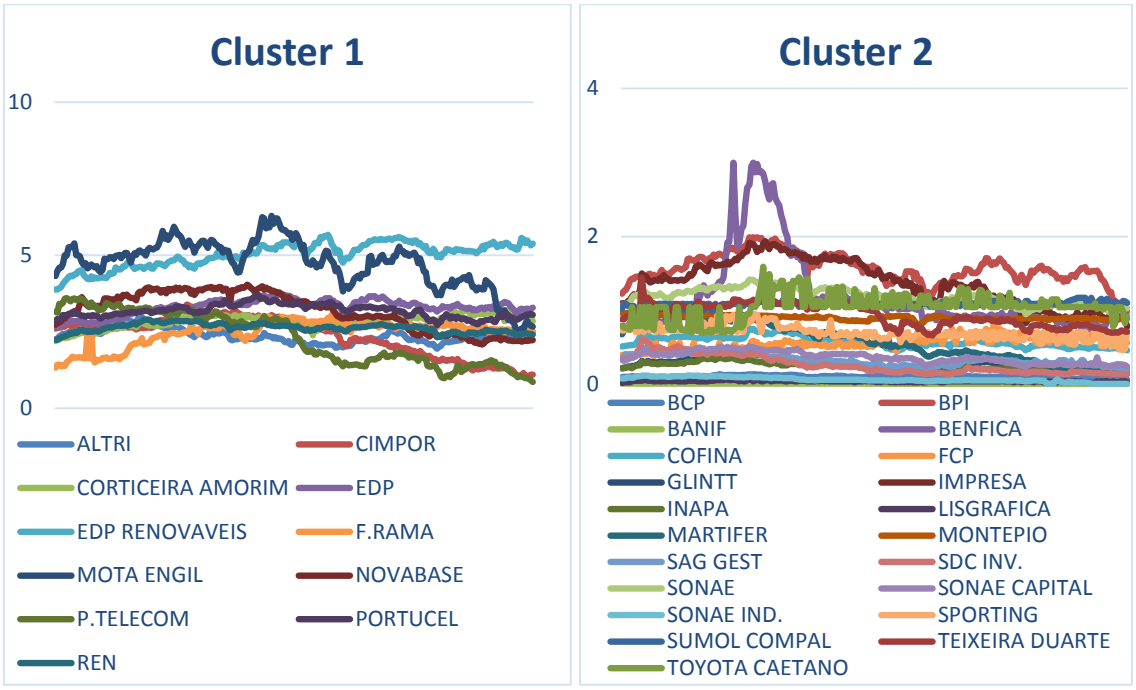


Figura 11 - Resultados dos clusters 1 e 2 do PAM com as distâncias euclidiana e Manhattan

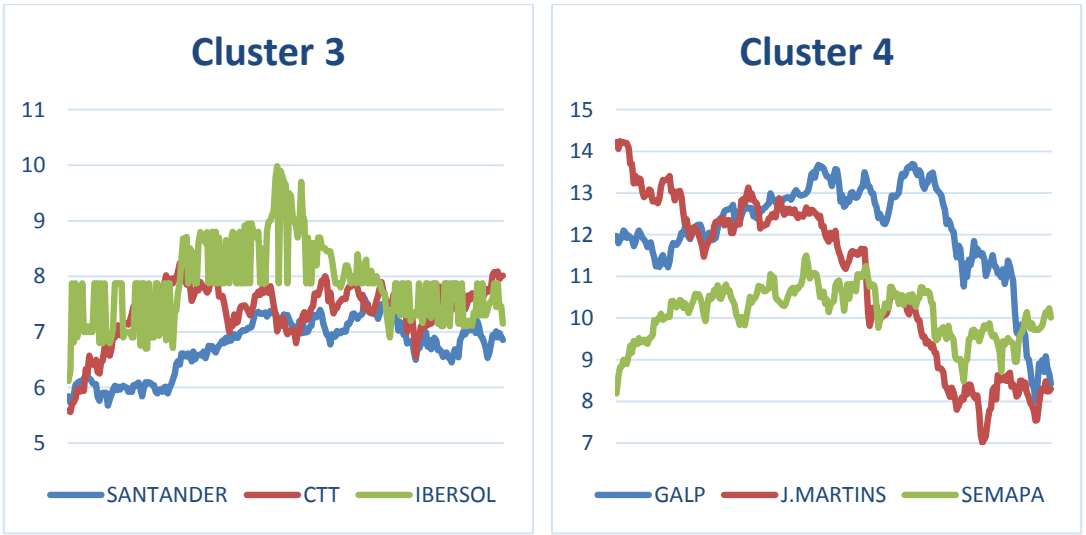


Figura 12 - Resultados dos clusters 3 e 4 do PAM com as distâncias euclidiana e Manhattan

Analisando a distribuição dos *clusters* com as duas distâncias verificamos que obtemos 4 *clusters* iguais. Podemos também verificar que o *cluster* 2 é constituído por 21 empresas, o que representa 58% das empresas, o que não é um bom resultado.

3.4.3. Método *Single Linkage*

Consideramos que o melhor número de *clusters* para o *Single Linkage* com a distância euclidiana foi de 10 *clusters*, uma vez que foi considerada a melhor partição para 62% dos índices e com a distância de *Manhattan* foi de 8 *clusters* uma vez que foi considerada a melhor partição para 54% dos índices.

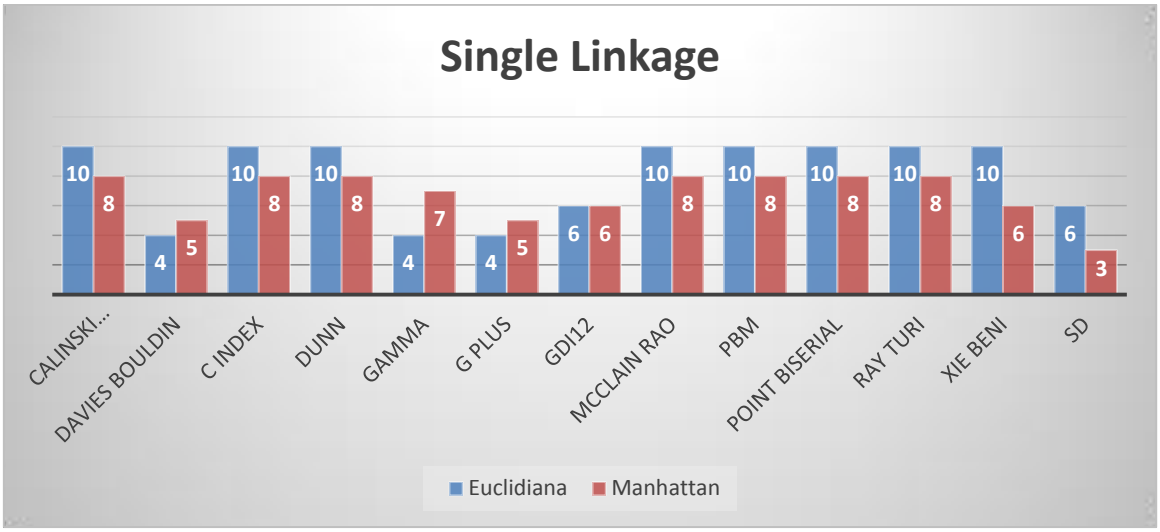


Figura 13 - Resultados dos índices pelo Single Linkage

Através dos índices *Calinski Harabasz, C index, Dunn, McClain Rao, PBM, Point Biserial, Ray Turi e Xie Beni* obtivemos o melhor número de *clusters* para o *Single Linkage* com a distância euclidiana. (10 clusters)

Através dos índices *Callinski Harabasz. C Index, Dunn, McClain Rao, PBM, Point Biserial* e *Ray Turi* obtivemos o melhor número de *clusters* para o *Single Linkage* com a distância euclidiana. (8 *clusters*)

Comparando os resultados das duas distâncias não obtivemos o melhor número de *clusters* igual com nenhum índice.

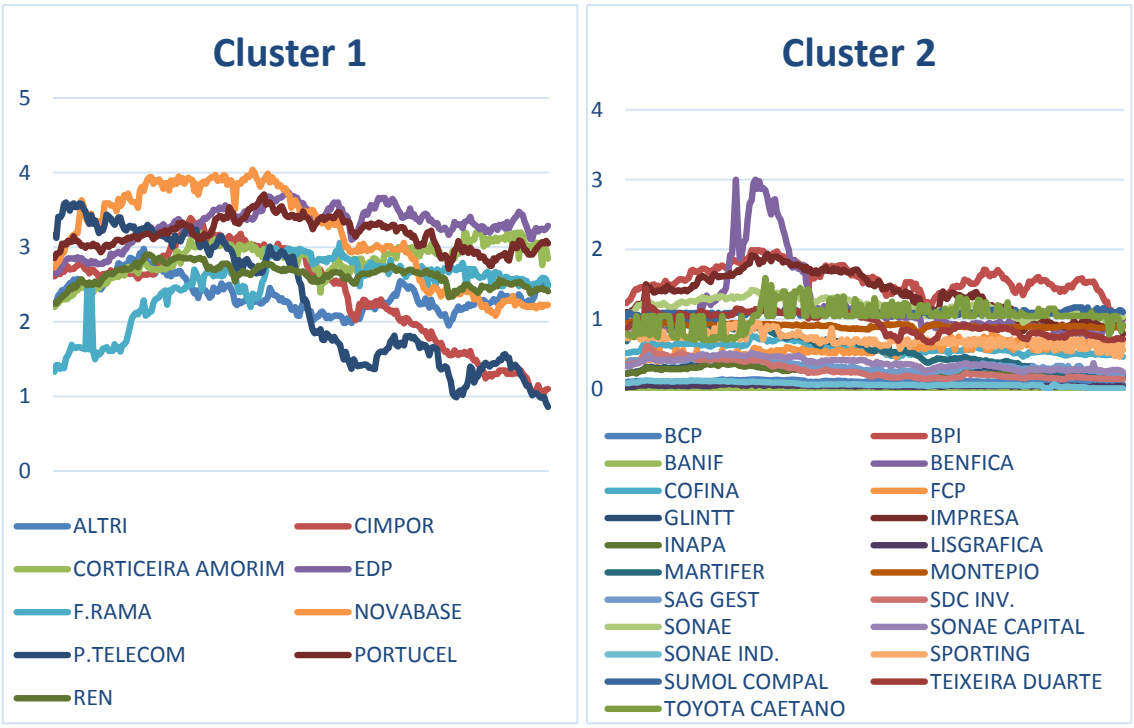


Figura 14 - Resultados dos clusters 1 e 2 do Single Linkage/Euclidiana

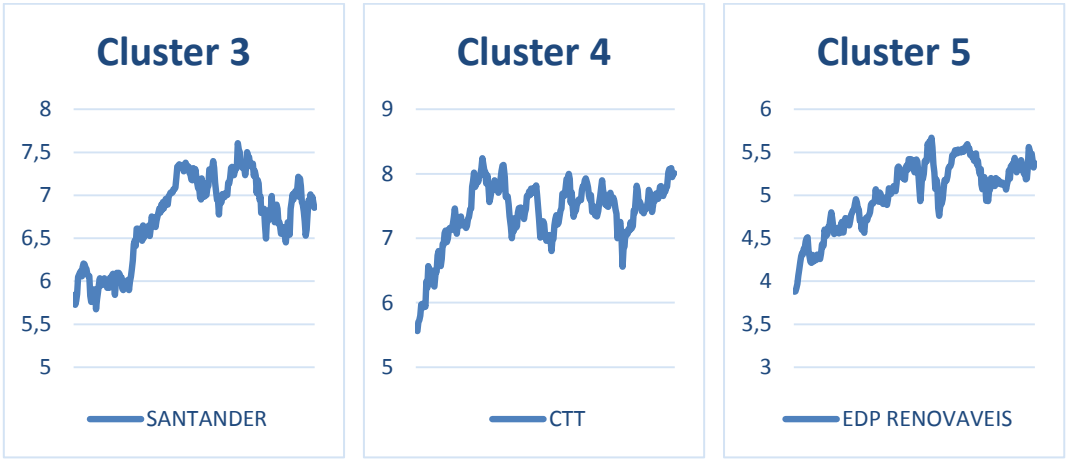


Figura 15 - Resultados dos clusters 3, 4 e 5 do Single Linkage/Euclidiana

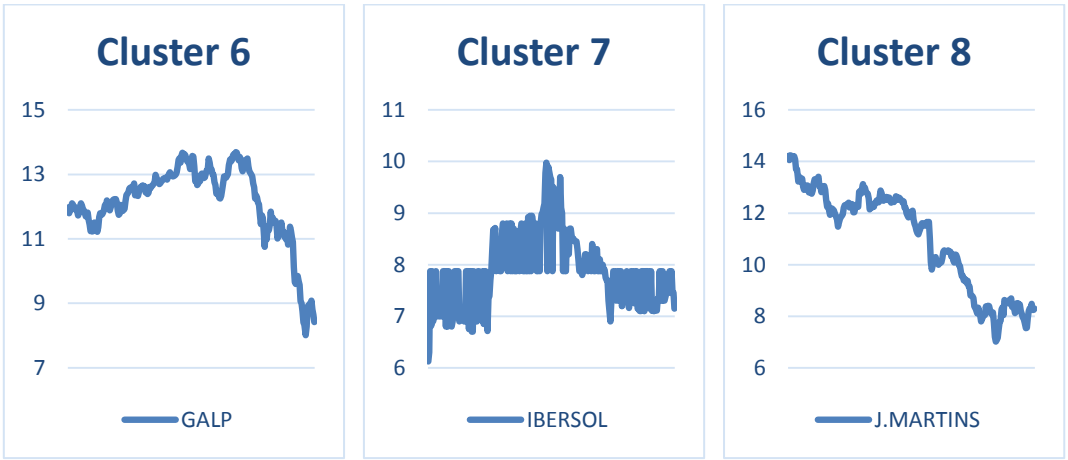


Figura 16 - Resultados dos clusters 6,7 e 8 do Single Linkage/Euclidiana

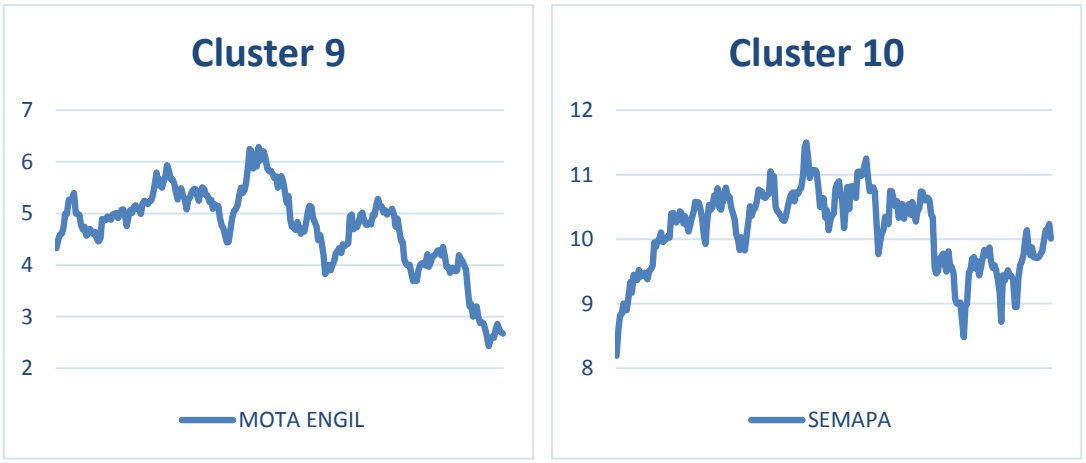


Figura 17 - Resultados dos clusters 9 e 10 do Single Linkage/Euclidiana

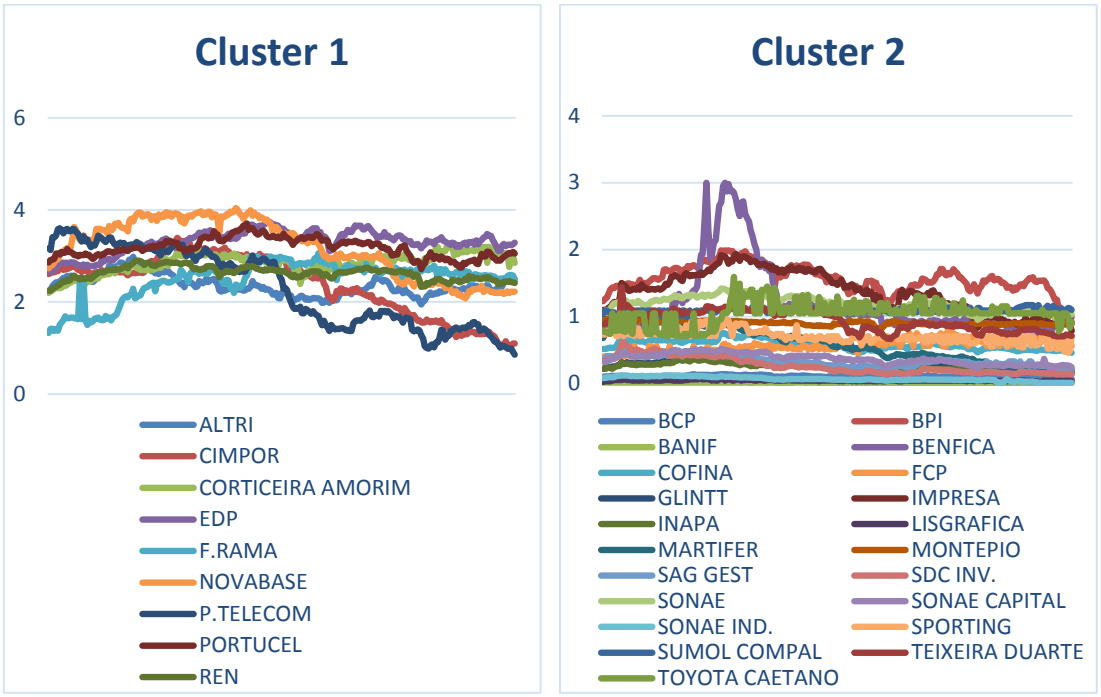


Figura 18 - Resultados dos clusters 1 e 2 do Single Linkage/Manhattan

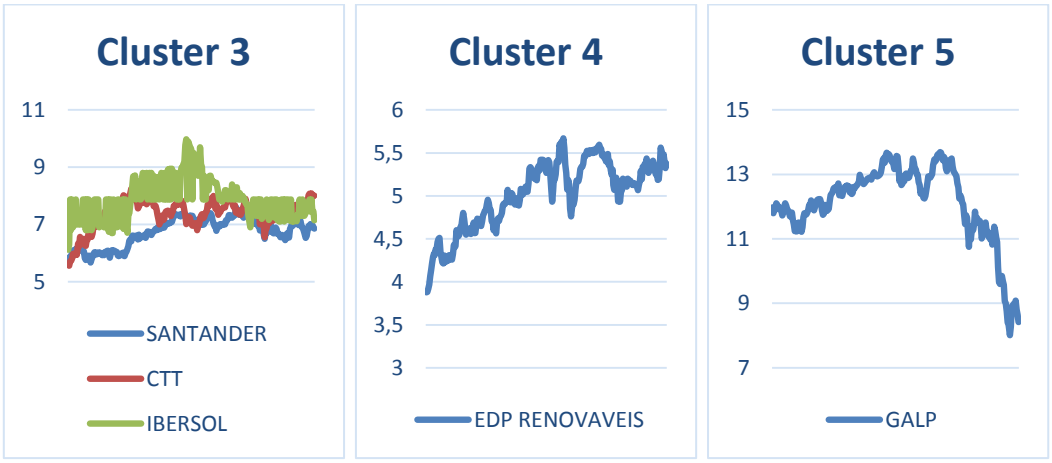


Figura 19 - Resultados dos clusters 3,4 e 5 do Single Linkage/Manhattan

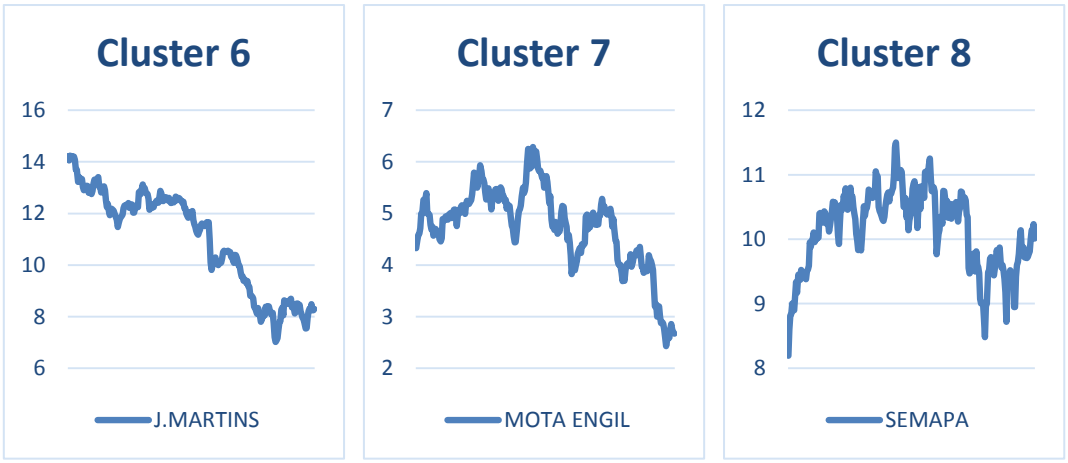


Figura 20 - Resultados dos clusters 6,7 e 8 do Single Linkage/Manhattan

Analisando a distribuição dos *clusters* podemos verificar que não obtivemos bons resultados com as duas distâncias. Com a distância euclidiana obtivemos oito *clusters* com uma empresa em cada um e dois *clusters* com 9 empresas e 21 empresas (55% das empresas), respetivamente. Com a distância *Manhattan* obtivemos cinco *clusters* com 1 empresa em cada um e dois *clusters* com 9 empresas e 21 empresas (55% das empresas), respetivamente.

3.4.4. Método *Complete Linkage*

Consideramos que o melhor número de *clusters* para o *complete Linkage* com a distância euclidiana foi de 7 *clusters*, uma vez que foi considerada a melhor partição para 54% dos índices e com a distância de *Manhattan* foi de 9 *clusters* uma vez que foi considerada a melhor partição para 38% dos índices.

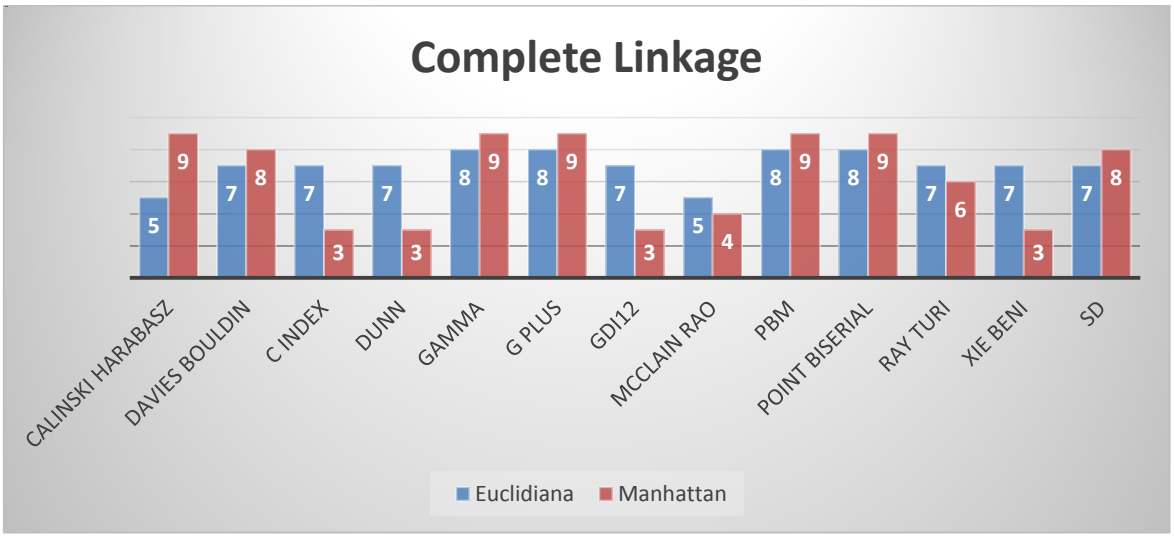


Figura 21 - Resultados dos índices pelo Complete Linkage

Através dos índices *Davies Bouldin,C index, Dunn, GD12, Ray Turi, Xie Beni* e *SD* obtivemos o melhor número de *clusters* para o *Complete Linkage* com a distância *euclidiana*. (7 *clusters*)

Através dos índices *Calinski Harabasz, Gamma, G Plus, PBM e Point Biserial* obtivemos o melhor número de *clusters* para o *Complete Linkage* com a distância *Manhattan*. (9 *clusters*)

Comparando os resultados das duas distâncias podemos verificar que obtivemos resultados diferentes com todos os índices.

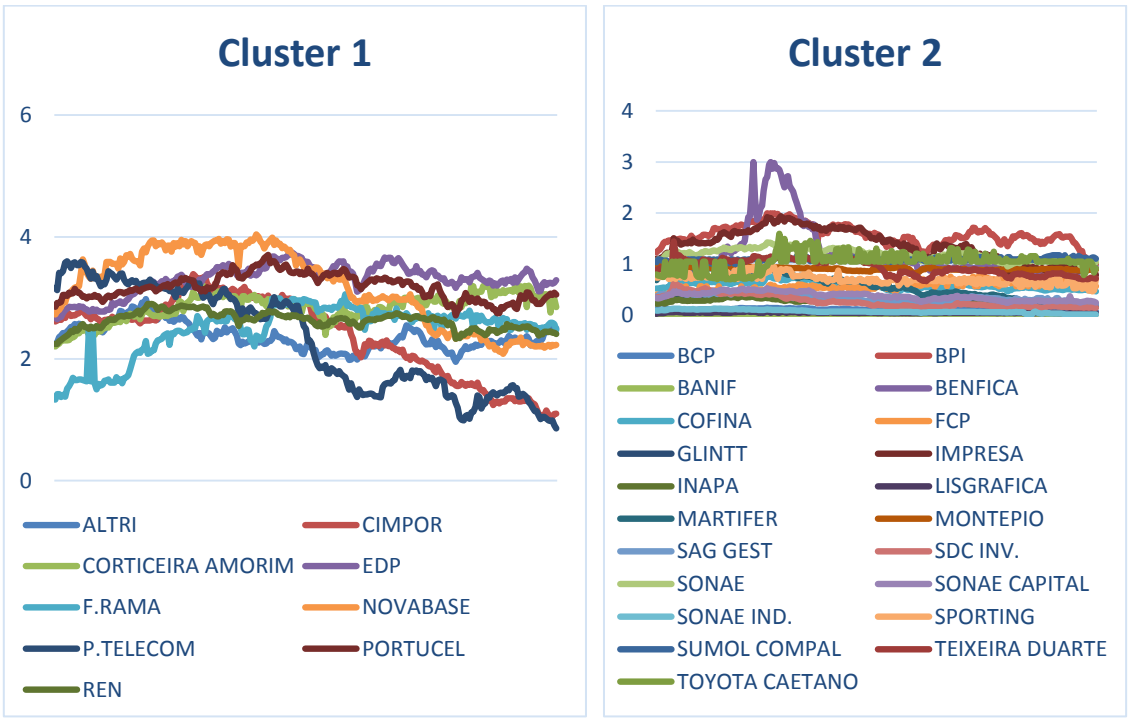


Figura 22 - Resultados dos clusters 1 e 2 do Complete Linkage/Euclidiana

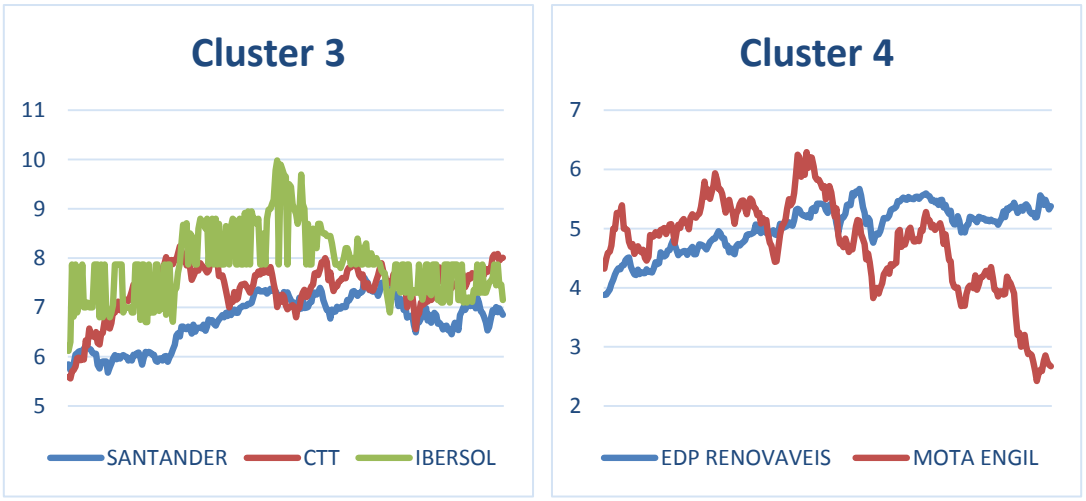


Figura 23 - Resultados dos clusters 3 e 4 do Complete Linkage/Euclidiana

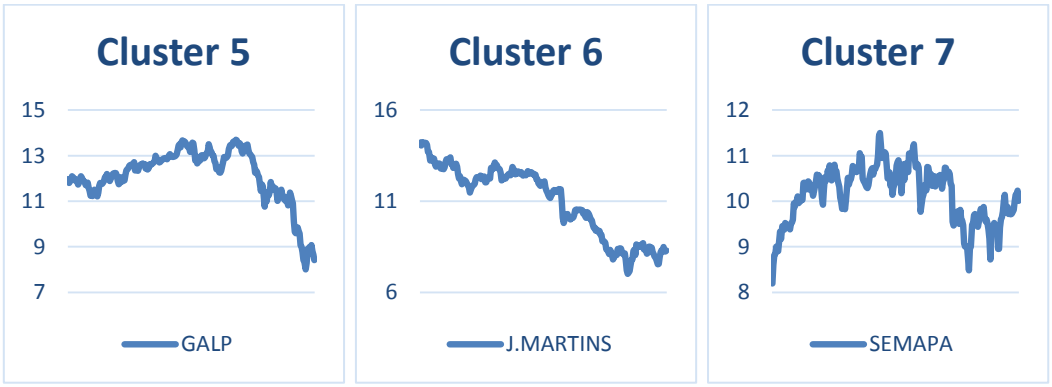


Figura 24 - Resultados dos clusters 5,6 e 7 do Complete Linkage/Euclidiana

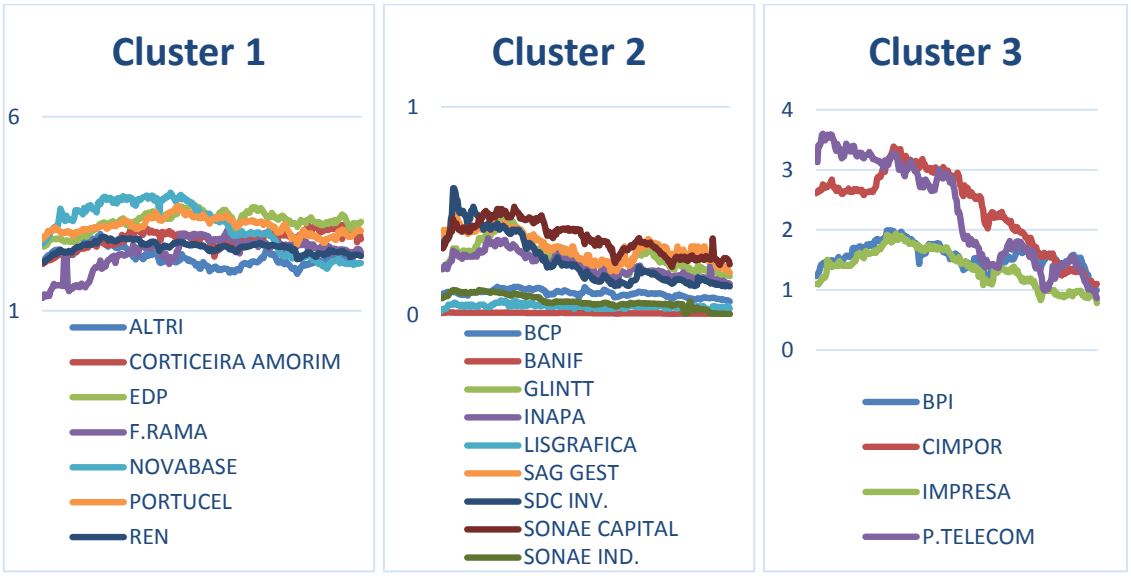


Figura 25 - Resultados dos clusters 1,2 e 3 do Complete Linkage/Manhattan

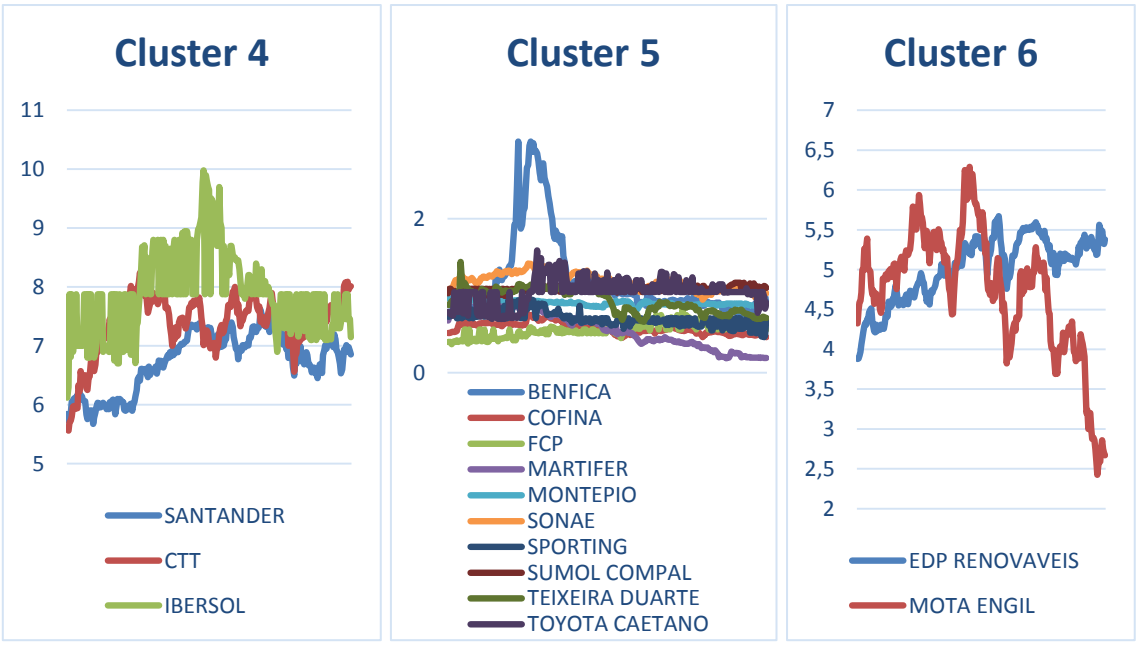


Figura 26 - Resultados dos clusters 4,5 e 6do Complete Linkage/Manhattan

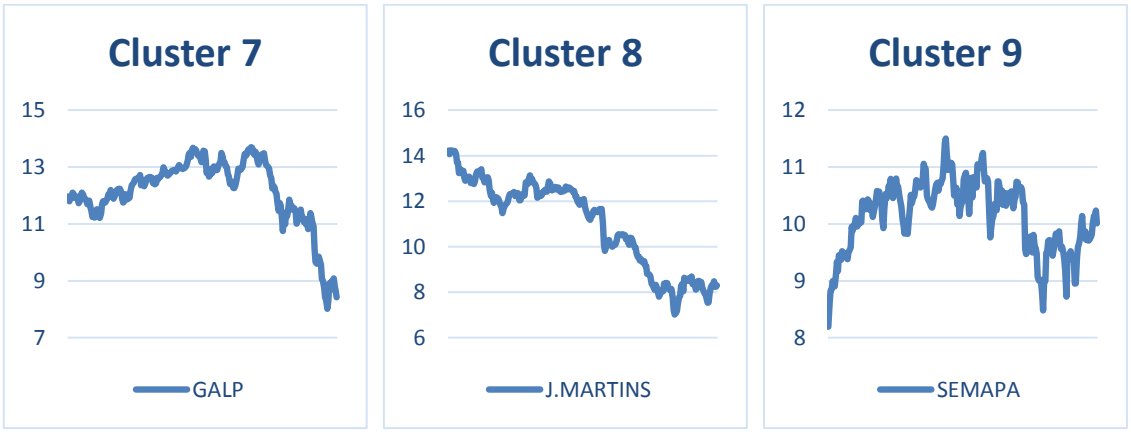


Figura 27 - Resultados dos clusters 7,8 e 8 do Complete Linkage/Manhattan

Analisando a distribuição das empresas pelos *clusters* com as duas distâncias verificamos que obtivemos cinco *clusters* iguais: três *clusters* com uma empresa em cada um (Galp, J.Martins e Semapa), um *cluster* com o Santander, CTT e a Ibersol e por fim um *cluster* com a EDP Renováveis e a Mota Engil.

3.4.5. Método *Average Linkage*

Consideramos que o melhor número de *clusters* para o A*verage Linkage* com a distância euclidiana foi de 11 *clusters*, uma vez que foi considerada a melhor partição para 46% dos índices e com a distância de *Manhattan* foi de 7 *clusters* uma vez que foi considerada a melhor partição para 38% dos índices.

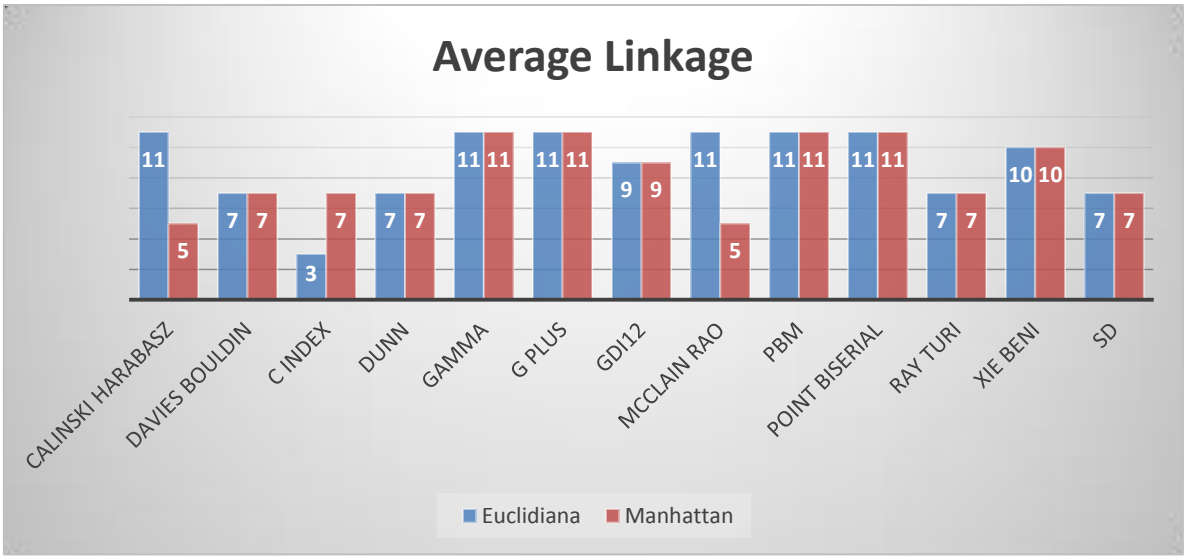


Figura 28 - Resultados dos índices pelo Average Linkage

Através dos índices *Calinski Harabasz, Gamma, G plus, McClain Rao, PBM e Point Biserial* obtivemos o melhor número de *clusters* para o *Average Linkage* com a distância euclidiana. (11 *clusters*)

Através dos índices *Davies Bouldin, C index, Dunn, Ray Turi e SD* obtivemos o melhor número de *clusters* para o *Average Linkage* com a distância *Manhattan*. (7 *clusters*)

Comparando os resultados das duas distâncias podemos verificar que com os índices *Calinlki Harabasz, C índex* e *McClain Rao* obtivemos o melhor número de *clusters* diferente.

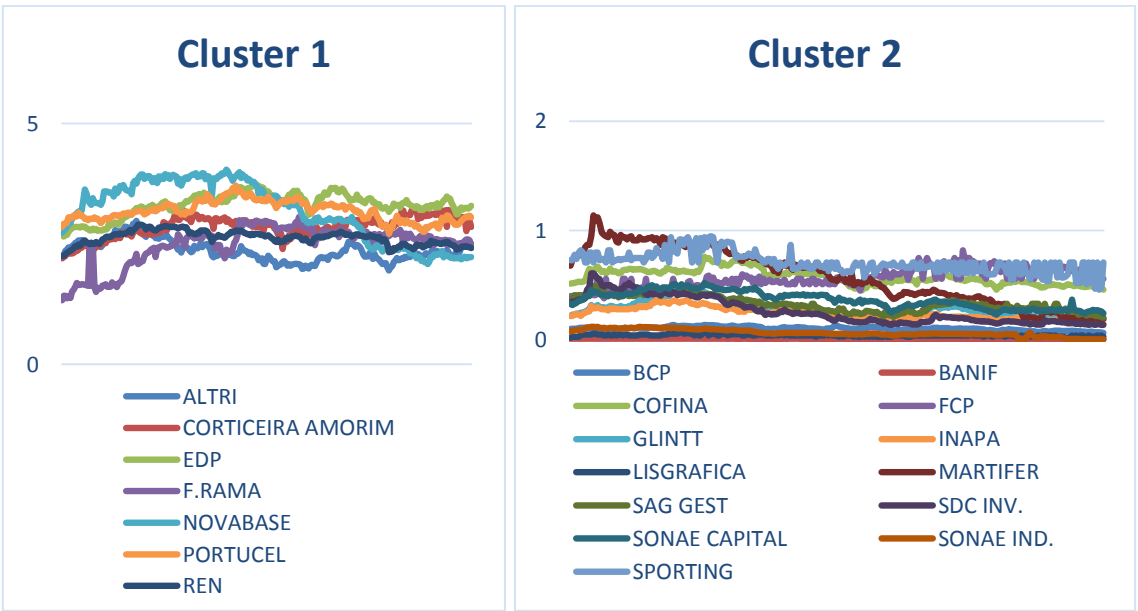


Figura 29 - Resultados dos clusters 1 e 2 do Average Linkage/Euclidiana

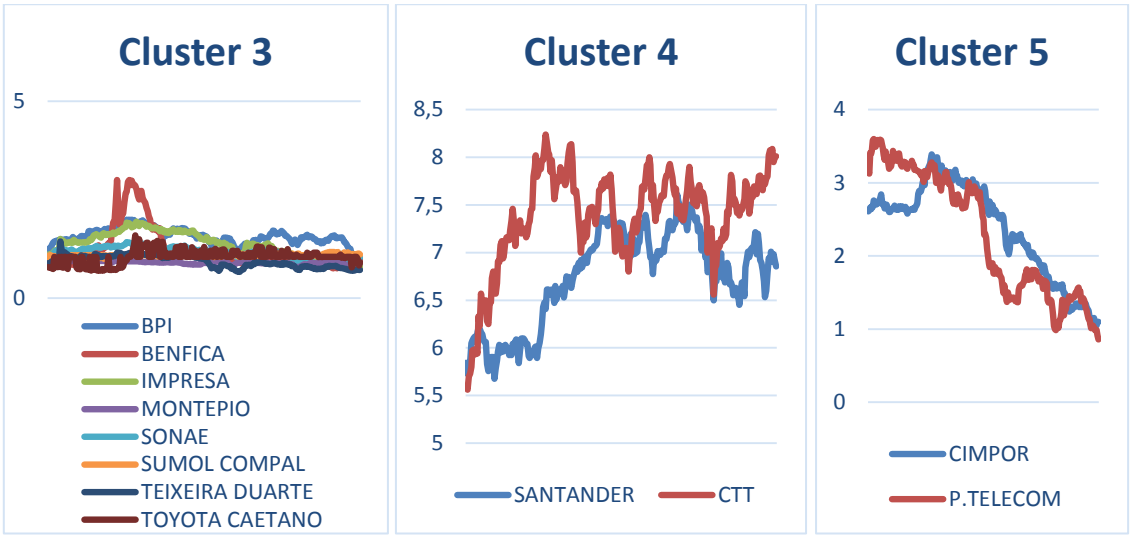


Figura 30 - Resultados dos clusters 3,4 e 5 do Average Linkage/Euclidiana

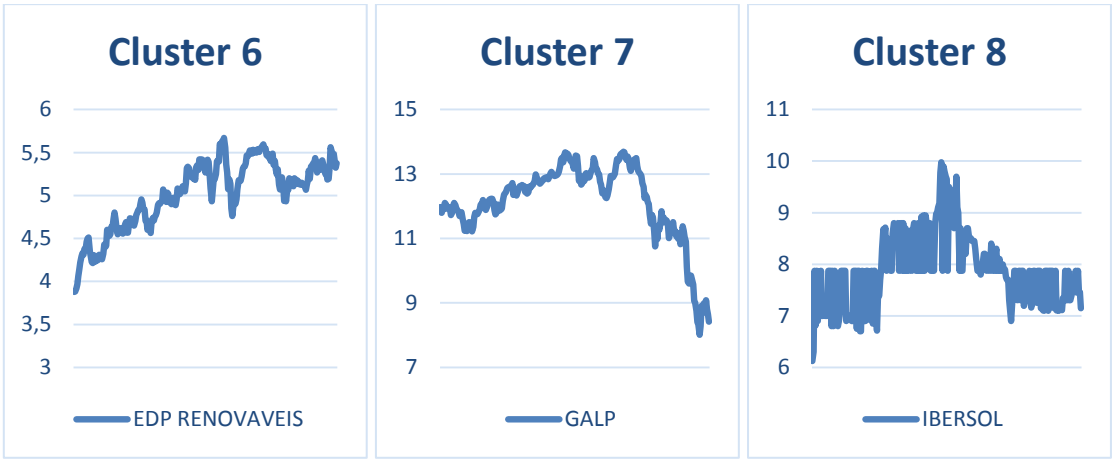


Figura 31 - Resultados dos clusters 6,7 e 8 do Average Linkage/Euclidiana

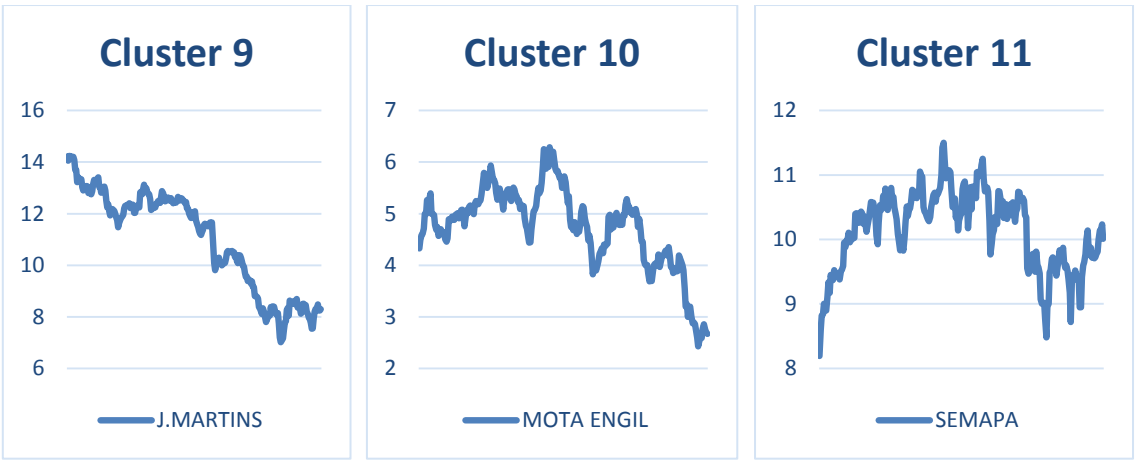


Figura 32 - Resultados dos clusters 9,10 e 11 do Average Linkage/Euclidiana

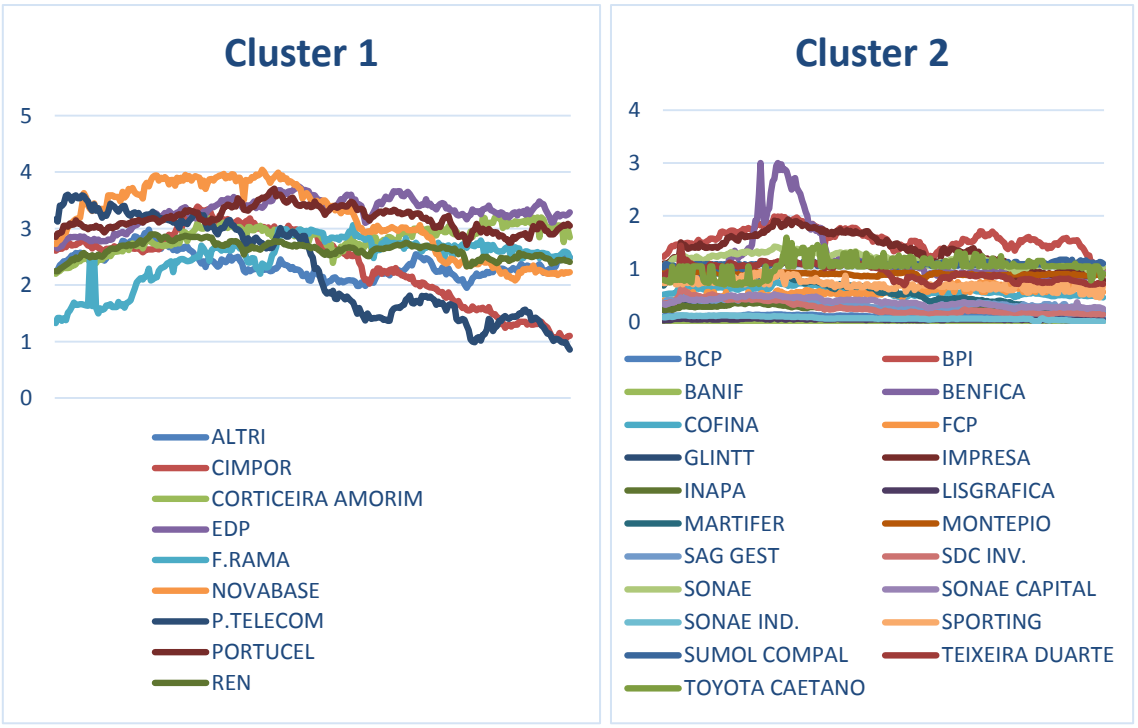


Figura 33 - Resultados dos clusters 1 e 2 do Average Linkage/Manhattan

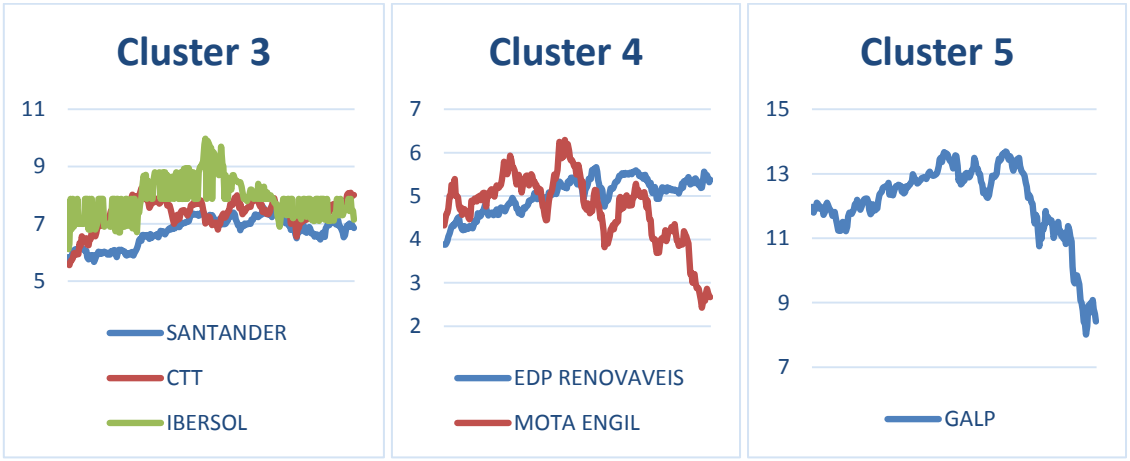


Figura 34 - Resultados dos clusters 3, 4 e 5 do Average Linkage/Manhattan

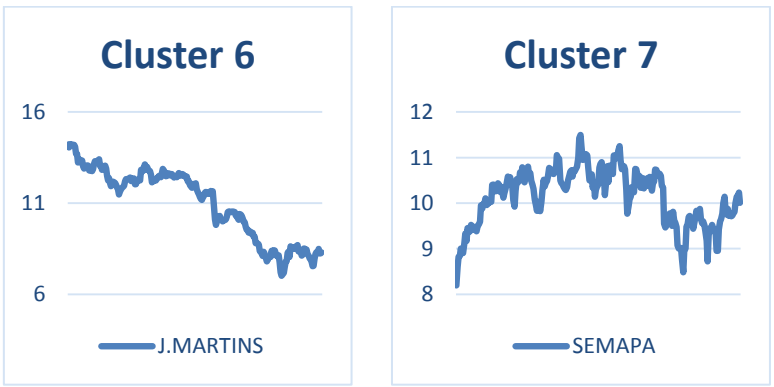


Figura 35 - Resultados dos clusters 6 e 7 do Average Linkage/Manhattan

Analisando a distribuição das empresas pelos *clusters* com as duas distâncias verificamos que três dos *clusters* são iguais: três *clusters* com uma empresa em cada um (Galp, J.Martins e Semapa).

Verificamos que com a distância euclidiana obtivemos 6 *clusters* com uma empresa em cada um e com a distância Manhattan obtivemos 3 *clusters* com uma empresa em cada um.

Por fim verificamos que um dos *clusters* que obtivemos com a distância de Manhattan é constituído por 21 empresas, ou seja, 55% das empresas, o que não é um bom resultado.

3.4.6. Diana

Consideramos que o melhor número de *clusters* para o *Diana* com a distância euclidiana foi de 7 *clusters*, uma vez que foi considerada a melhor partição para 58% dos índices e com a distância de *Manhattan* foi de 7 *clusters* uma vez que foi considerada a melhor partição para 54% dos índices.

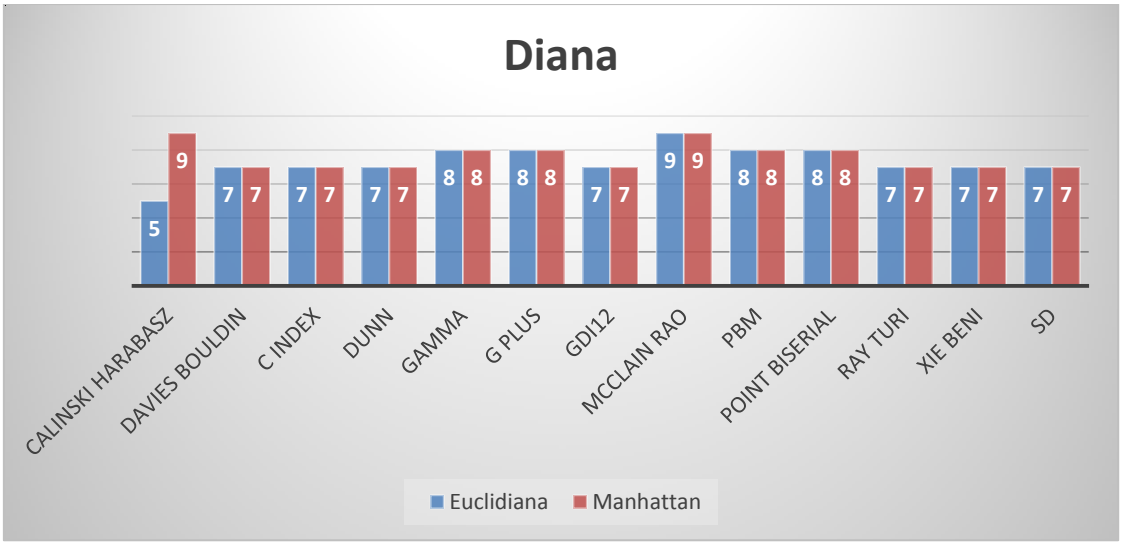


Figura 36 - Resultados dos índices pelo Diana

Através dos índices *Davies Bouldin, C index, Dunn, GDI12, Ray Turi, Xie Berie e SD* obtivemos o melhor número de clusters para o *Diana* com as duas distâncias. (7 *clusters*)

Comparando os resultados das duas distâncias apenas com o *Calinski Harabasz* é que obtivemos um melhor número de *clusters* diferente.

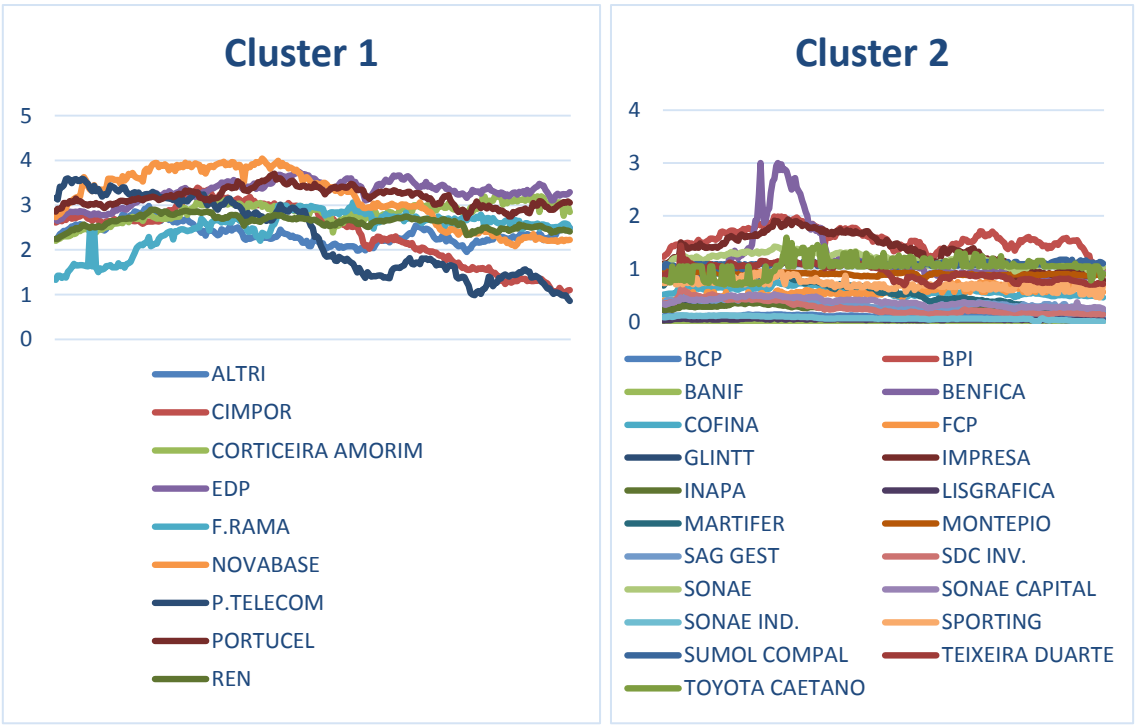


Figura 37 - Resultados dos clusters 1 e 2 do Diana tanto com a distância euclidiana como com a Manhattan

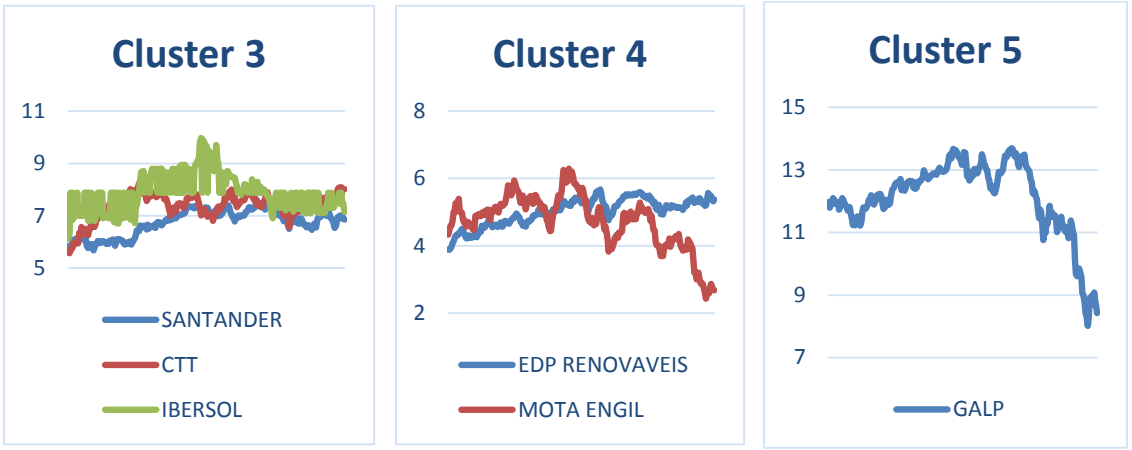


Figura 38 - Resultados dos clusters 3, 4 e 5 do Diana tanto com a distância euclidiana como com a Manhattan

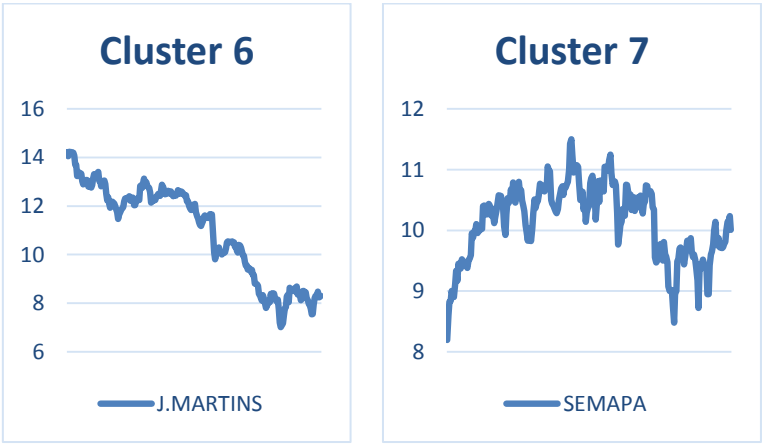


Figura 39 - Resultados dos clusters 6 e 7 do Diana tanto com a distância euclidiana como com a Manhattan

Analisando a distribuição das empresas pelos *clusters* verificamos que obtivemos os mesmos *clusters* com as duas distâncias.
Verificamos que obtivemos três *clusters* com uma empresa em cada *cluster* e um *cluster* com 21 empresas (55% das empresas) o que não é um bom resultado.

3.4.7. C-Means

Consideramos que o melhor número de *clusters* para o *C-Means* com a distância euclidiana foi de 7 *clusters*, uma vez que foi considerada a melhor partição para 31% dos índices e com a distância de *Manhattan* foi de 11 *clusters* uma vez que foi considerada a melhor partição para 23% dos índices.

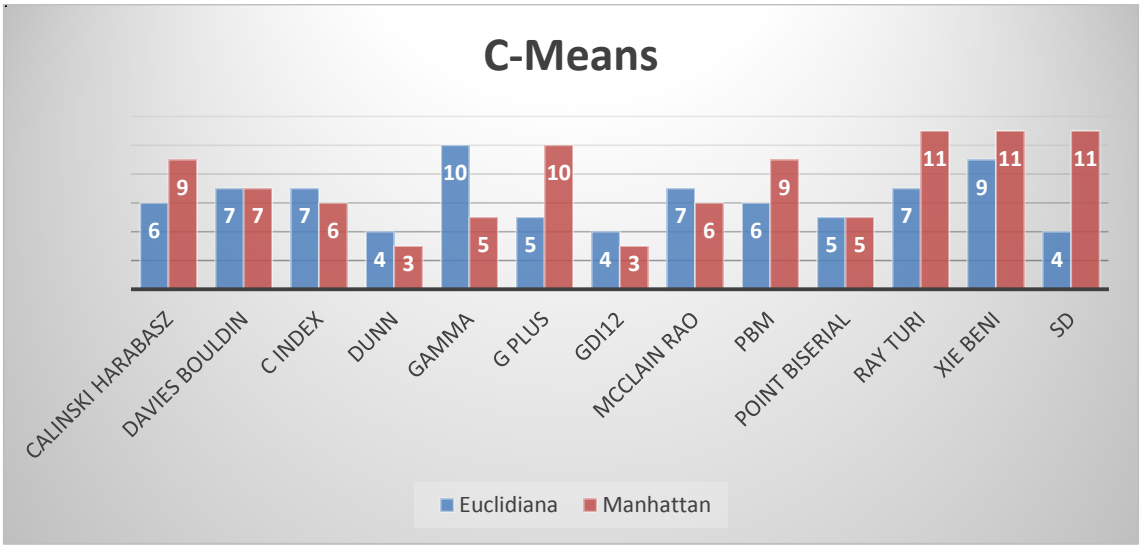


Figura 40 - Resultados dos índices pelo C-Means

Através dos índices *Davies Bouldin, C index, Mcclain Rao* e o *Ray Turi* obtivemos o melhor número de *clusters* para o *C-Means* com a distância euclidiana. (7 clusters)
Através dos índices *Ray Turi, Xie Beni e SD* obtivemos o melhor número de *clusters* para o *C-Means* com a distância *Manhattan*. (11 clusters)

Comparando os resultados com as duas distâncias obtivemos os mesmos resultados apenas com os índices *Davies Bouldin e Point Biserial.*

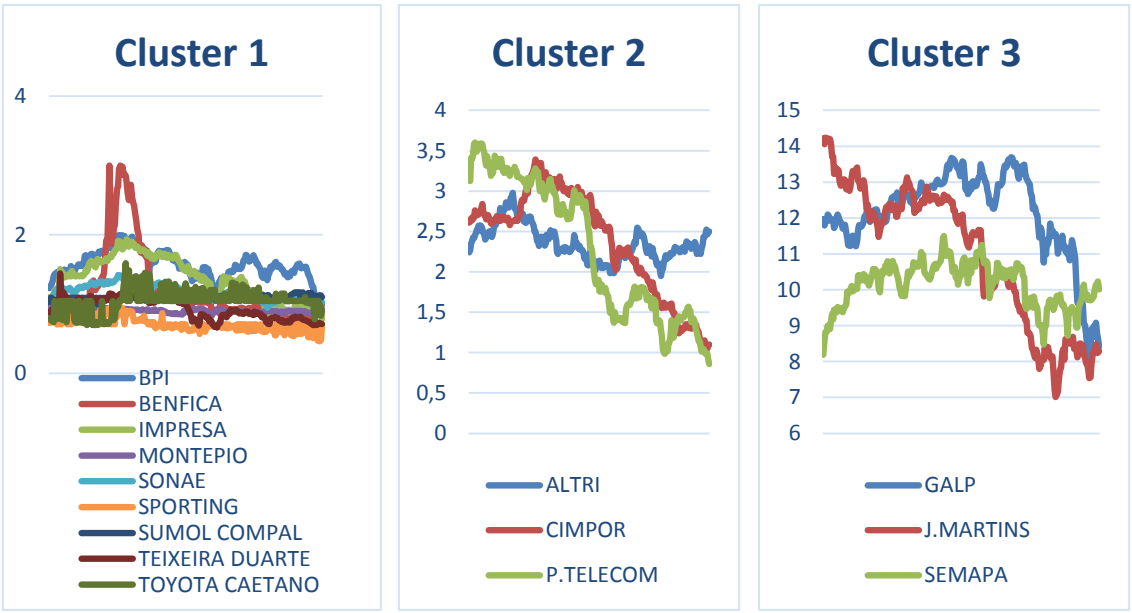


Figura 41 - Resultados dos clusters 1, 2 e 3 do C-Means/Euclidiana

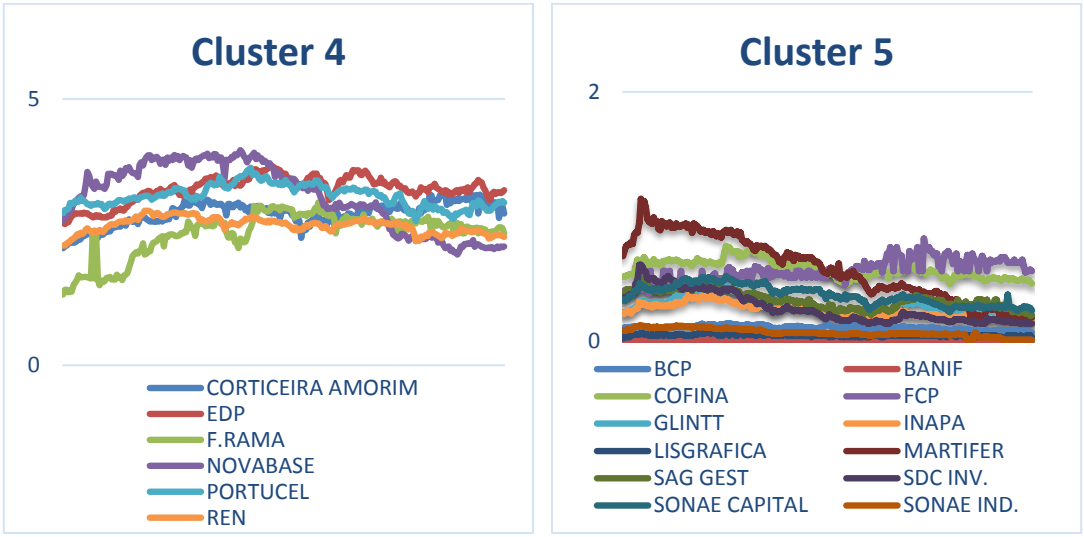


Figura 42 - Resultados dos clusters 4 e 5 do C-Means/Euclidiana

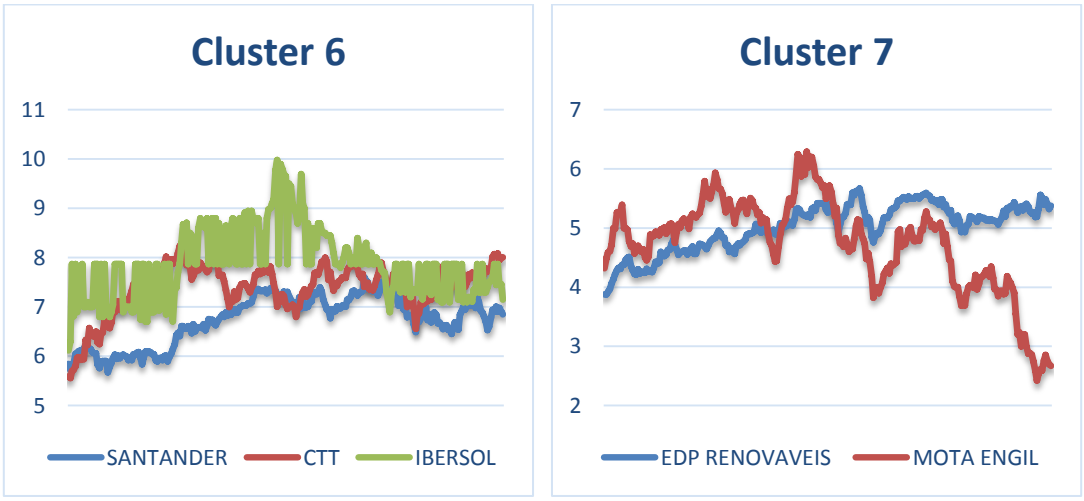


Figura 43 - Resultados dos clusters 6 e 7 do C-Means/Euclidiana

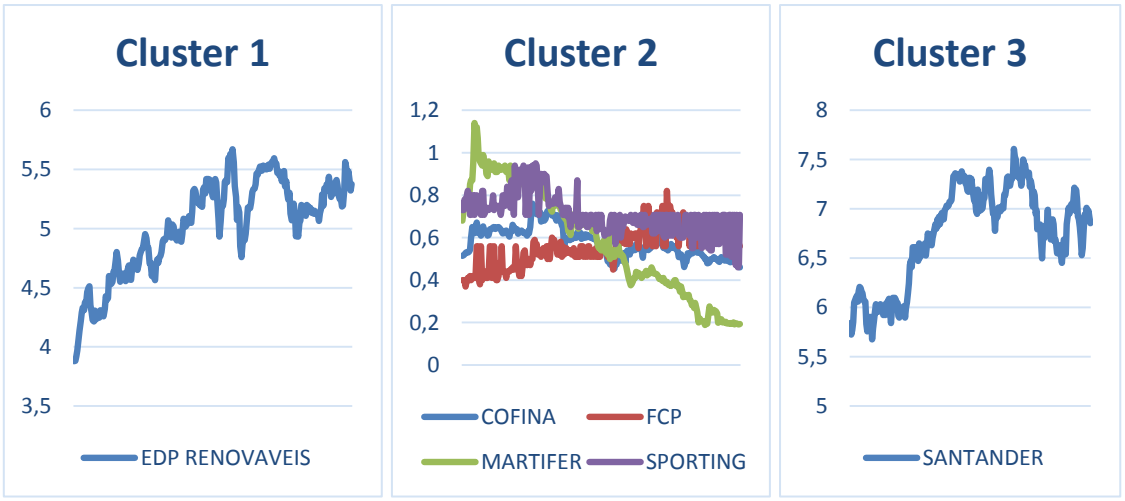


Figura 44 - Resultados dos clusters 1,2 e 3 do C-Means/Manhattan

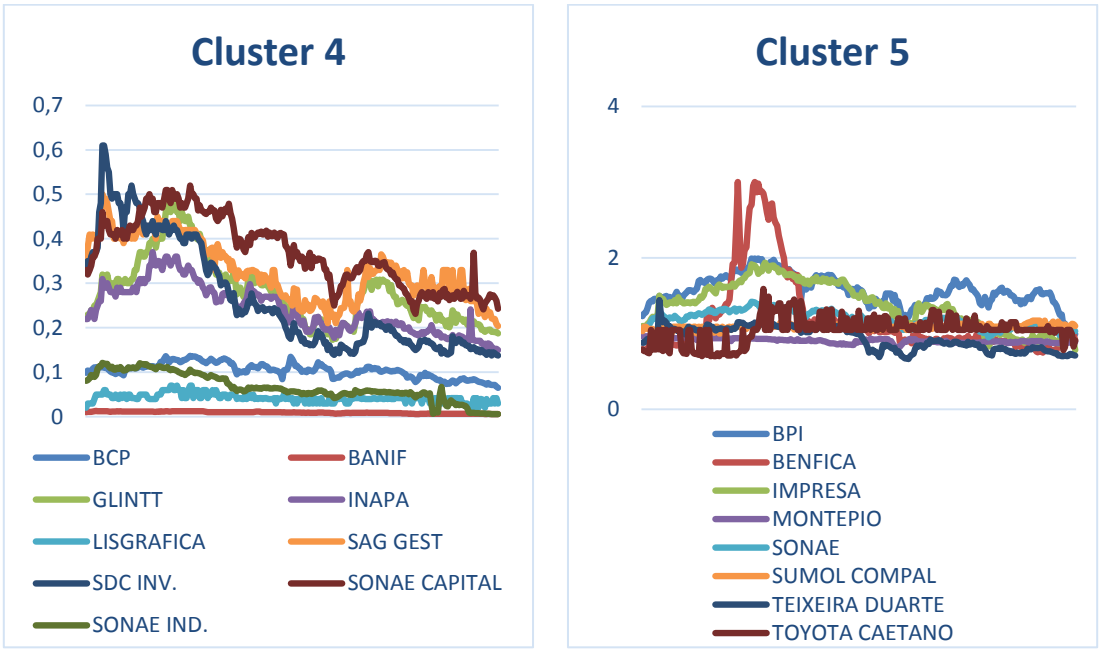


Figura 45 - Resultados dos clusters 4,5 do C-Means/Manhattan

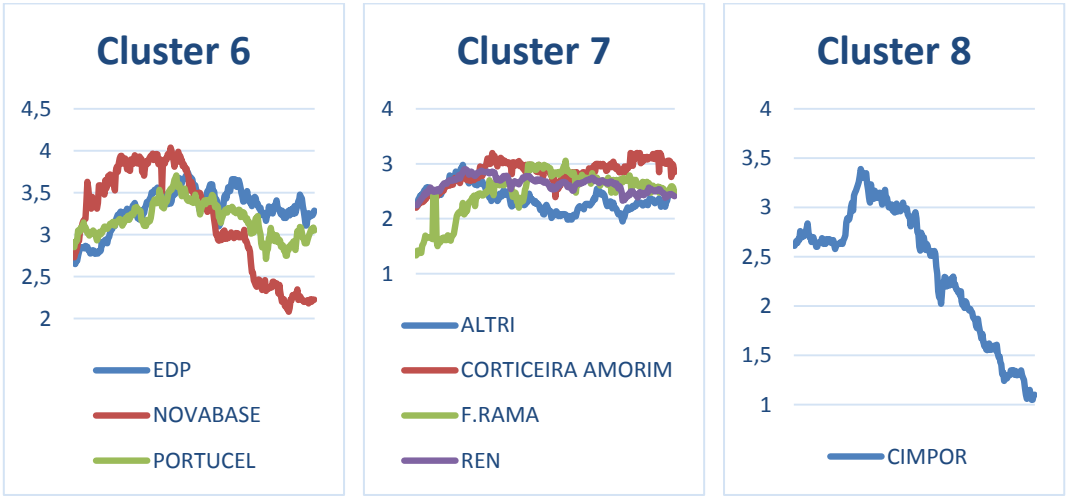


Figura 46 - Resultados dos clusters 6,7 e 8 do C-Means/Manhattan

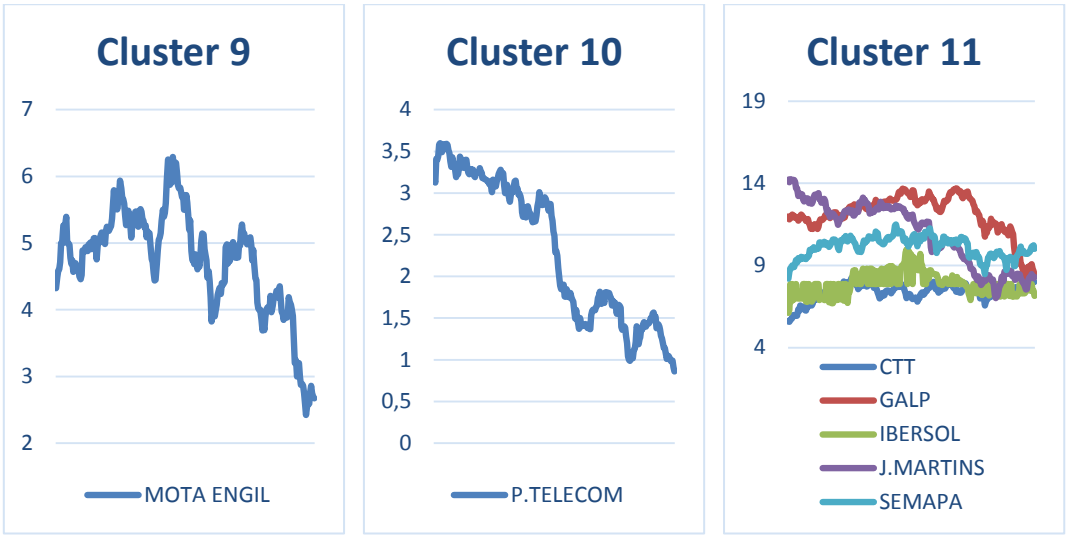


Figura 47 - Resultados dos clusters 9,10 e 11 do C-Means/Manhattan

Analisando a distribuição das empresas pelos *clusters* verificamos que obtivemos *clusters* diferentes com as duas distâncias.

Com a distância *Manhattan* obtivemos cinco *clusters* com uma empresa em cada um e com a distância euclidiana não obtivemos nenhum *clusters* com apenas uma empresa.

3.4.8. Funny

Consideramos que o melhor número de *clusters* para *Funny* com a distância euclidiana foi de 8 *clusters*, uma vez que foi considerada a melhor partição para 54% dos índices e com a distância de *Manhattan* foi de 8 *clusters* uma vez que foi considerada a melhor partição para 31% dos índices.

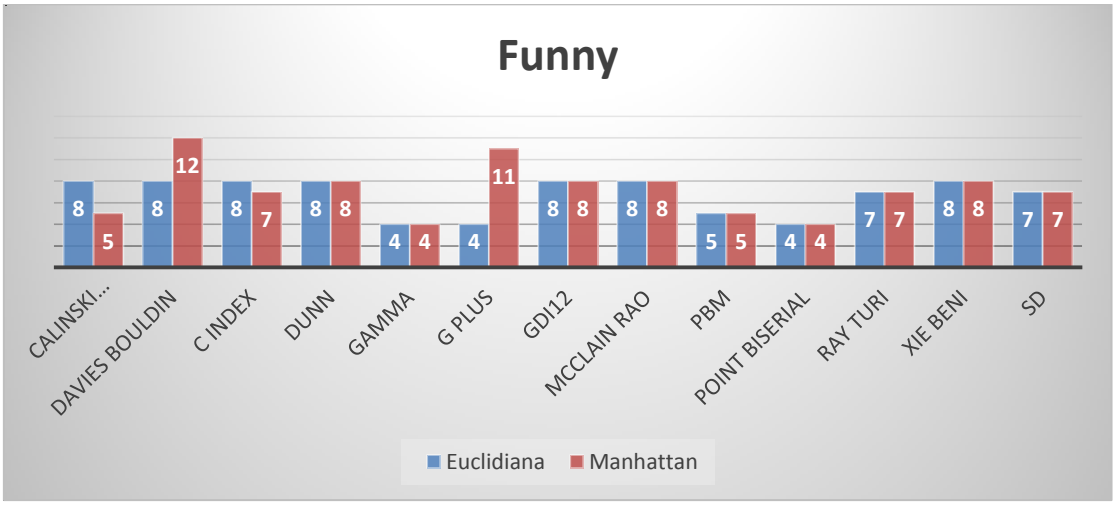


Figura 48- Resultados dos índices pelo Funny

Através dos índices *Calinski Harabasz, Davies Bouldin, C index, Dunn, GDI12, McClain Rao* e *Xie Beni* obtivemos o melhor número de *clusters* para o *Funny* com a distância euclidiana. (8 clusters)

Através dos índices *Dunn, GDI12, McClain Rao* e *Xie Beni* obtivemos o melhor número de *clusters* para o *Funny* com a distância *Manhattan*. (8 *clusters*)

Comparando os resultados com as duas distâncias obtivemos resultados diferentes com os índices *Calinski Harabasz, Davies Bouldin, C índex, Gamma e G Plus.*

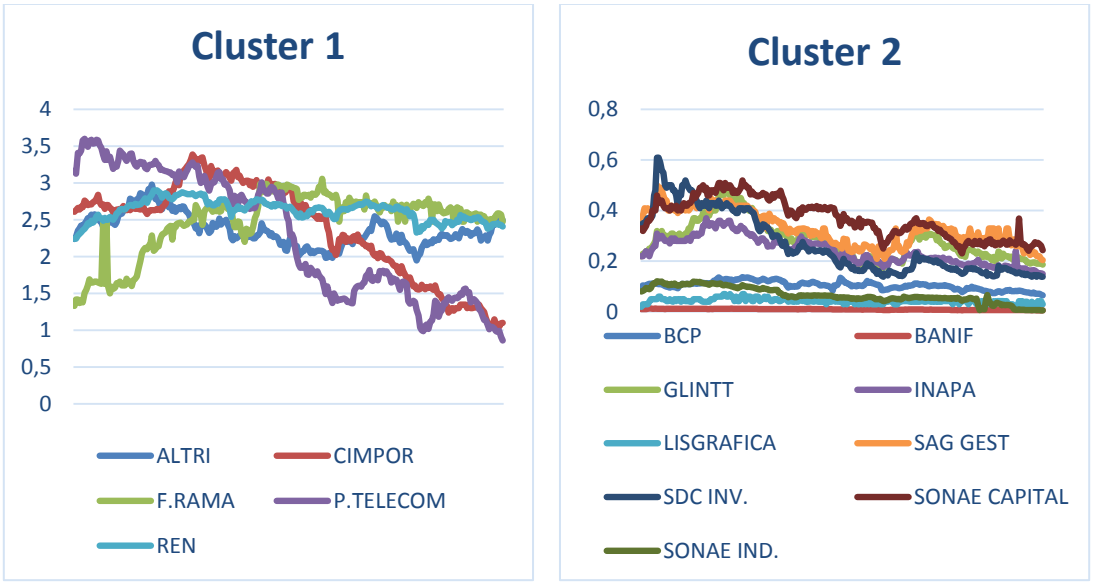


Figura 49 - Resultados dos clusters 1 e 2 do Funny/Euclidiana

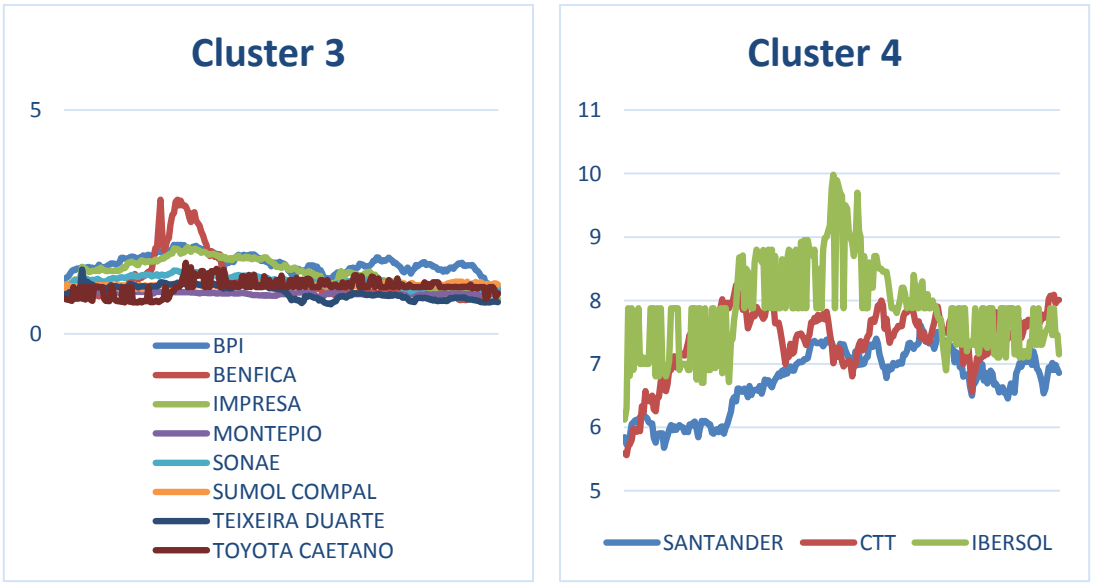


Figura 50 - Resultados dos clusters 3 e 4 do Funny/Euclidiana

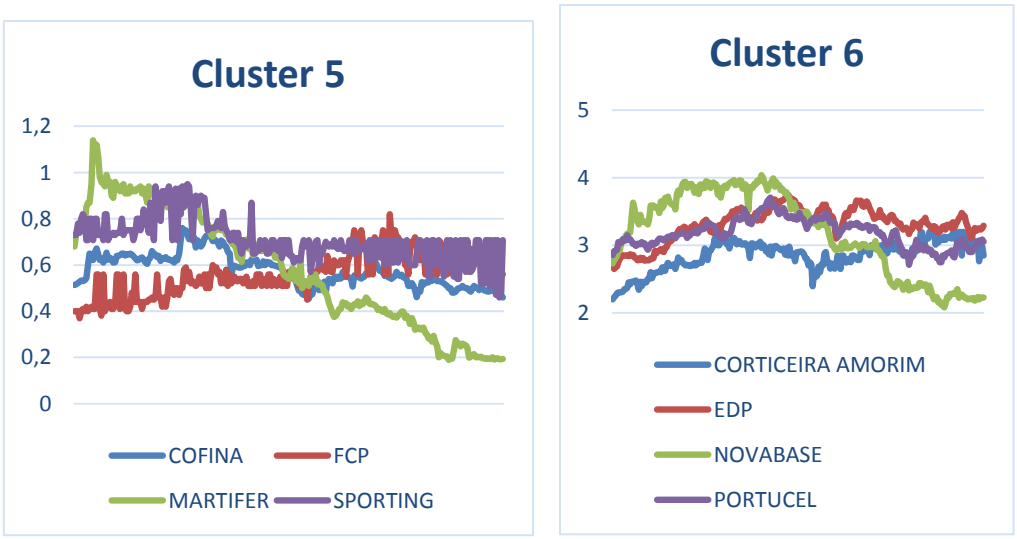


Figura 51 - Resultados dos clusters 5 e 6 do Funny/Euclidiana

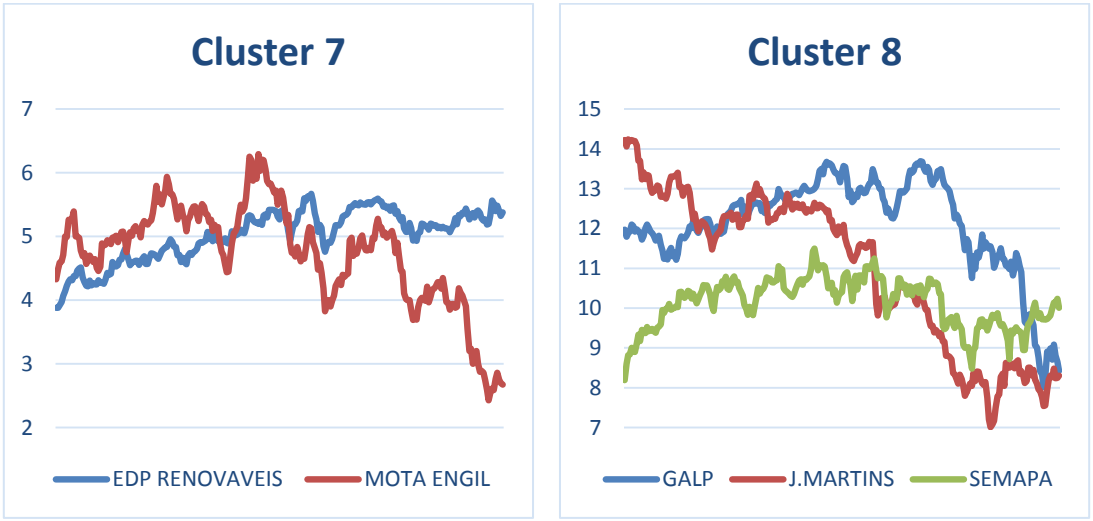


Figura 52 - Resultados dos clusters 7 e 8 do Funny/Euclidiana

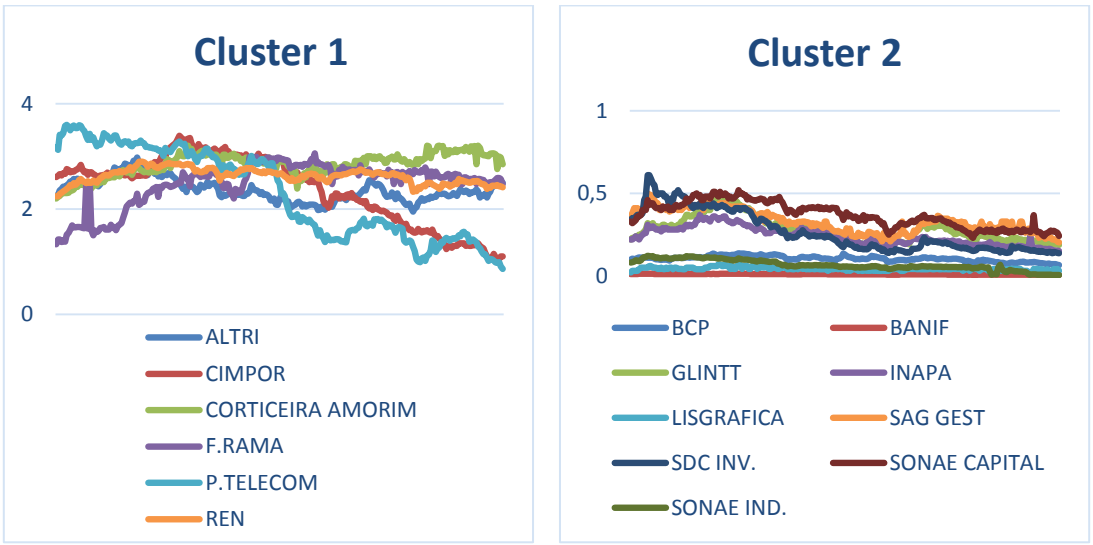


Figura 53 - Resultados dos clusters 1 e 2 do Funny/Manhattan

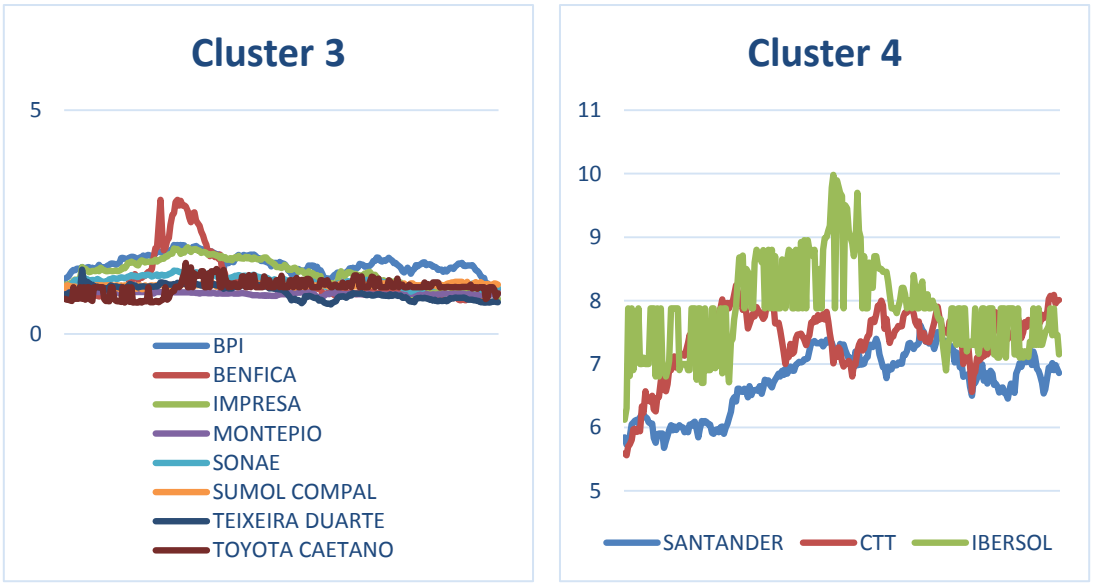


Figura 54 - Resultados dos clusters 3 e 4 do Funny/Manhattan

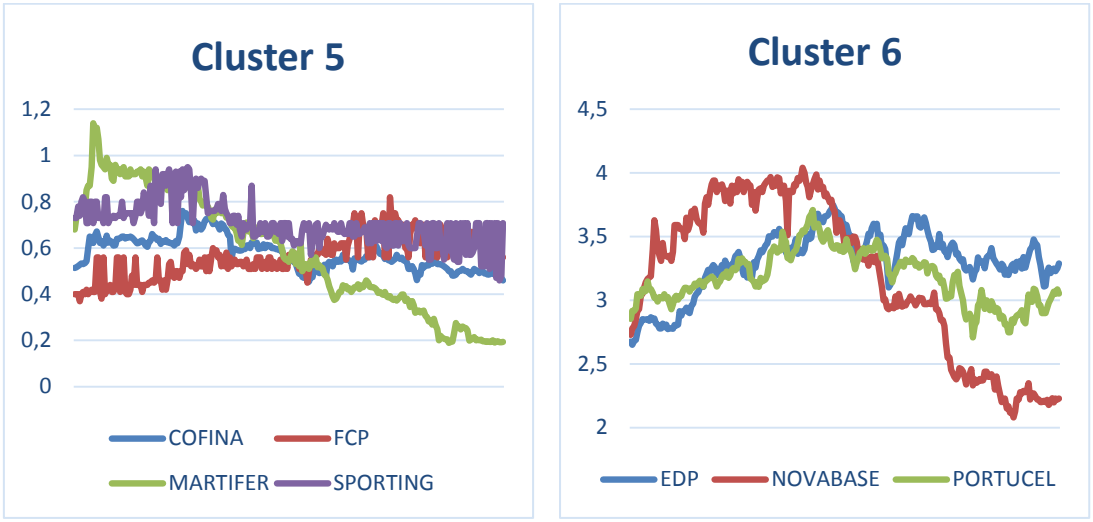


Figura 55 - Resultados dos clusters 5 e 6 do Funny/Manhattan

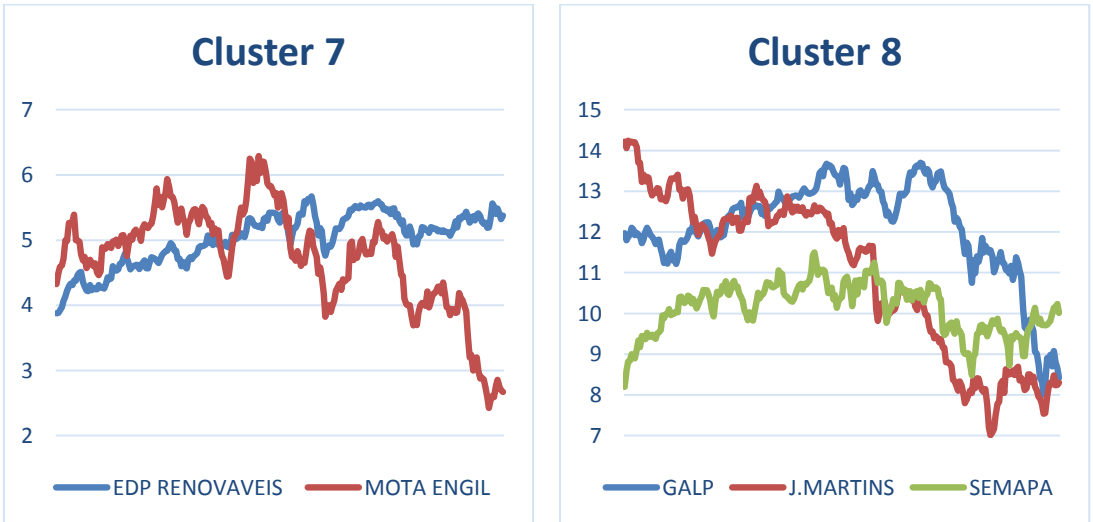


Figura 56 - Resultados dos clusters 7 e 8 do Funny/Manhattan

Podemos verificar que na distribuição das empresas pelos *clusters* a única diferença entre os resultados com as duas distâncias é a localização da empresa Corticeira Amorim. Nos resultados obtidos com a distância euclidiana esta empresa aparece com a EDP, Novabase e Portucel e nos resultados obtidos com a distância *Manhattan* aparece com a Altri, Cimpor, F.Rama, P.Telecom e Ren.

3.5. Conclusões

Analisando todos os resultados de todos os métodos/ medidas o melhor número de *clusters* foi de 7 clusters obtidos em 31% dos métodos/medidas. Os métodos que obtivemos 7 *clusters* foram o *Complete Linkage/Euclidiana, Average Linkage/Manhattan Diana/Euclidiana, Diana/Manhattan e C-Means/Euclidiana*.

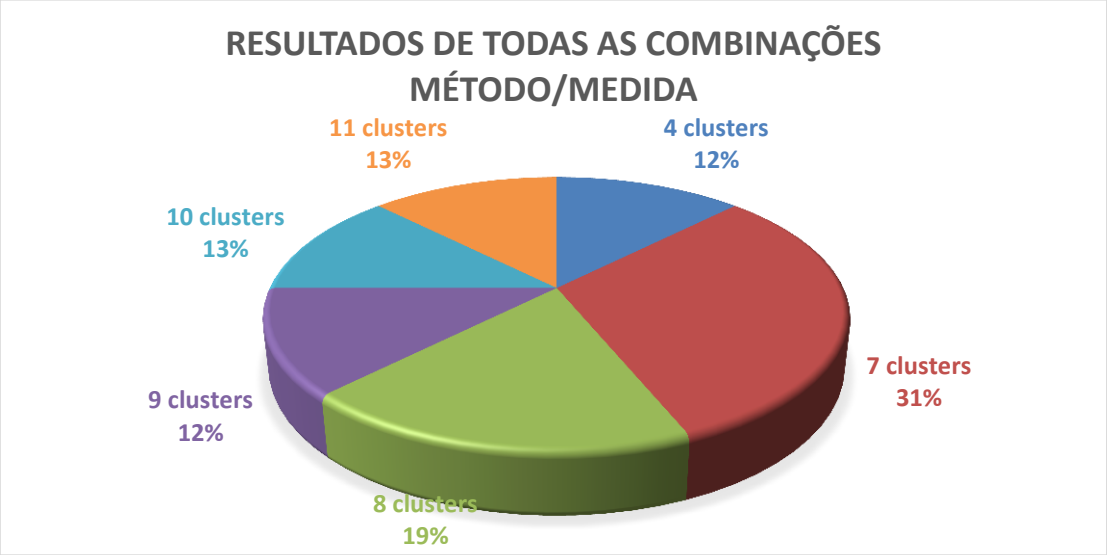


Figura 57 - Resultados de todas as combinações método/medida

Tabela 3 - Percentagem de índices que obtivemos 7 clusters como o melhor número

Percentagem de índices que obtivemos 7 clusters como o melhor número	**Método/Medida**
54%	*Complete Linkage*/Euclidiana
38%	*Average Linkage/Manhattan*
54%	*Diana*/Euclidiana
54%	*Diana/Manhattan*
31%	*C-Means*/Euclidiana

Nos métodos *Complete/Eucidiana, Average/Manhattan, Diana/Euclidiana Diana/Manhattan, PAM/euclidiana, PAM/Manhattan, Single Linkage/Euclidiana e Single Linkage/Manhattan* obtivemos um *cluster* com 21 empresas (55% das empresas). Assim podemos concluir que os algoritmos com piores resultados, nesta dissertação, foram o *PAM, Single Linkage* e o *Diana*. Relativamente ao *Single Linkage* era um resultado esperado uma vez que é conhecido que dos métodos hierárquicos este algoritmo normalmente é o algoritmo que se obtém piores resultados em comparação com o *Complete Linkage* e *Average Linkage.*

Pela aplicação dos diferentes métodos verificámos que a grande maioria dos casos, as empresas Santander, CTT e Ibersol foram colocadas num mesmo *cluster*. Relativamente às empresas Galp, J.Martins e Semapa ou foram colocadas juntas num mesmo *cluster*, ou foram consideradas *outliers*, uma vez que foram colocadas sozinhas num cluster. Também verificamos que para a maioria dos casos a EDP Renováveis e a Mota Engil forma colocados num mesmo *cluster* ou são consideradas *outliers*.

Analisando as empresas por sectores verificamos que apenas a Galp, que é a única empresa do sector *Oil & Gas* foi colocada num *cluster* sozinha em alguns métodos ao contrário da P.Telecom, que é a única empresa do sector *Telecomunications*, que não foi colocada em nenhum *cluster* sozinho. Assim comparando os resultados dos vários algoritmos/métodos podemos concluir que os *clusters* não são definidos pelos sectores a que a empresa pertence. Os setores a que cada empresa pertence estão no anexo B.

4. Conclusão e trabalhos futuros

Como seria de esperar a principal conclusão desta dissertação é que o mercado de ações é um problema muito complexo e que mesmo existindo diversos estudos nesta área as conclusões continuam a ser muito subjetivas.

Nesta dissertação focamo-nos nos algoritmos de *Clustering* e nos índices para avaliar o melhor número de *clusters* e por isso este estudo pode ser considerado um primeiro passo na otimização de portfólios. O passo seguinte seria analisar quais seriam as empresas a escolher na otimização de portfólios, sendo que essas empresas pertenceriam a *clusters* diferentes.

Em relação ao facto de alguns autores [7,11,8] referirem que as empresas do mesmo sector "movem-se juntas" ao longo do tempo e fazer sentido uma vez que os fatores que fazem com que o preço de uma empresa descer/subir também vai fazer descer/subir o preço das empresas do mesmo sector, isso não implica necessariamente que outras empresas de outros sectores também não desçam/subam o preço devido a esses fatores. Mais uma vez é uma ideia subjetiva e por vezes não acontece.

Seria interessante num trabalho futuro aplicar mais algoritmos como o *SOM, DBSCAN, OPTICS, CLARA, AGNES, BIRCH, CURE e ROCK* ao conjunto de dados estudados nesta dissertação assim como utilizar mais distâncias como por exemplo a distância *Minkowski* ou a distância *Canberra*.

Bibliografia

[1] Mahajan K. S. and Kulkarni R.V, A review: Application of Data Mining tools for stock Market (2013).

[2] Hajizadeh E., Ardakani H. D. and Shahrabi J., Application of data mining techniques in stock markets: A survey (2010).

[3] Cai F., Le-khac N.A. and Kechadi M. T., Clustering Approaches for Financial Data Analysis: a Survey.

[4] Sherdiwala K. B., Data Mining Techniques in Stock Market (2014).

[5] Nanda S.R., Mahanty B. and Tiwari M.K., Clustering Indian stock market data for portfolio management (2010).

[6] Basalto N., Bellotti R., De Carlo F., Facchi P. and Pascazio S., Hausdorff clustering of financial time series (2005).

[7] Musetti A. T. Y, Clustering for financial time series (2012).

[8] Gavrilov M., Anguelov D., Indyk P. and Motwani R., Mining the Stock Market: which Measure is Best? (2000)

[9] Doherty K.A.J, Adams R.G., Davey N. and Pensuwon W., Hierarchical Topological Clustering Learns Stock Market Sectors (2005).

[10] Fallahpur S., Zadeh M.H. and Lakvan E.N., Use of Clustering Approach for Portfolio Management (2014).

[11] Wittman T., Time-Series Clustering and Association Analysis of Financial Data (2002).

[12] Moldovan D. and Silaghi G.C., Gene Trajectory Clustering for Learning the Stock Market Sectors (2009).

[13] Liao T.W., Clustering of Time Series data – a survey (2005).

[14] Michaud P., Clustering techniques (1997).

[15] Kovács F., Legány C. and Babos A., Cluster Validity Measurement Techniques (2005).

[16] Halkidi M., Batistakis Y. and Vazirgiannis M., On Clustering Validation Techniques (2001).

[17] Han J, Kamber M., Pei J. and Fraser S., Data Mining Concepts and Techniques Third Edition (2012).

[18] Milligan G. W. and Cooper M.C., Methodology review clustering methods (1987).

[19] Maimon O and Rokach L. Data Mining and Knowledge Discovery Handbook

[20] Xu R., Wunsch D.,Clustering (2009).

[21] Everitt B.S., Landau S., Leese M. and Stahl D., Cluster Analysis (2010).

[22] Bezdek, J.C., Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell, MA, USA (1981).

[23] Dunn, J.C., A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. J. Cybernet, Vol. 3, pp. 32–57 (1973).

[24] Wang W. and Zhang Y., On fuzzy cluster validity indices (2007).

[25] Klawonn F., Fuzzy Clustering: Insights and a new approach (2004).

[26] Wang C.J., Jin C., Fang H., Moormann A. and Wang H., A New Integrated Fuzzifier Evaluation and Selection (NIFEs) Algorithm for Fuzzy Clustering (2015).

[27] Suganya R. and Shanthi R. , Fuzzy C- Means Algorithm- A Review (2012).

[28] Rousseeuw P.J. and Kaufman L., Finding Groups in Data: An Introduction to Cluster Analysis (2008).

[29] Rousseeuw P., Hubert M. and Struyf A., Clustering in an Object-Oriented Environment (1997).

[30] Rendón E., Abundez I., Arizmendi A. and Quiroz E. M., Internal versus External cluster validation indexes (2011).

[31] Dimitriadou E., Dolnicar S. and Weingessel A., An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets (1999).

[32] Calinski T. and Harabasz J., A dendrite method for cluster analysis, Communications in Statistics, 3, 1–27 (1974).

[33] Tomasev N. and Radovanović M., Clustering Evaluation in High-Dimensional Data

[34] Putler D. S. and Krider R. E., Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R (2012).

[35] McClain J. O. and Rao V. R., Clustisz: A program to test for the quality of clustering of a set of objects. Journal of Marketing Research, 12, 456–460 (1975).

[36] Milligan G.W. and Cooper M.C., An Examination Procedures for determining the number of clusters in a data set (1985).

[37] Desgraupes B., Clustering Indices (2013).

[38] Charrad M., Ghazzali N., Boiteau V. and Niknafs A., NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set (2014).

[39] Hubert, L. J. and Levin, J. R. A general statistical framework for assessing categorical clustering in free recall. Phycological Bulletin, 83, 1072-1080 (1976).

[40] Guerra L., Robles V., Bielza C. and Larrañaga P., A comparison of clustering quality índices using outliers and noise (2012).

[41] Baker, F. and Hubert L. Measuring the power of hierarchical cluster analysis. Journal of the American Statistical Association 31–38 (1975).

[42] Charrad M., Lechevallier Y., Ahmed M.B. and Saporta G., On the Number of Clusters in Block Clustering Algorithms (2010).

[43] Rohlf F. J., Methods of comparing classifications. Annual Review of Ecology and Systematics, 5:101-113 (1974).

[44] S. Ray and Rose H. Turi., Determination of number of clusters in k-means clustering and application in colour image segmentation. in Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, 137-143 (1999).

[45] Bandyopadhyay S. Pakhira M. K. and Maulik U. Validity index for crisp and fuzzy clusters. Pattern Recognition, 37:487-501 (2004).

[46] Chakrabarty A. and Purkayastha B.S., A Kernel PBM Index for Fuzzy Clustering of Numeric Data (2010).

[47] Davies D. L. and Bouldin D. W. , A cluster Separation Measure (1979).

[48] Yeh J.-H. and Joung F.-J., Lin J.-C., CDV Index: A Validity Index for Better Clustering Quality Measurement (2014).

[49] Mirkin B., Clustering for Data Mining: A Data Recovery Approach (2013).

[50] X.L. Xie and G. Beni. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(4):841-846, (1991).

[51] Dunn J., Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics, 4:95-104 (1974).

[52] Halkidi M., Batistakis Y. and Vazirgiannis M., Clustering Validity Checking Methods: Part II (2002).

[53] Bezdek , J. C. and Pal N. R. , Some New Indexes of Cluster Validity (1998).

[54] Halkidi, M., Vazirgiannis, M. and Batistakis, I., Quality scheme assessment in the clustering process, Proceedings of PKDD, Lyon, France, (2000).

[55] http://www.bolsadelisboa.com.pt/cotacoes/accoes-lisboa

[56] Silva H. B., Brito P. and Pinto da Costa J., A partitional clustering algorithm validated by a clustering tendency index based on graph theory. Pattern Recognition, Volume 39, Issue 5, pages 776-788 (2006).

Anexos

Esta dissertação é constituída por dois anexos: Anexo A e Anexo B. O anexo A é constituído por todos os resultados de todos os índices aplicados a vários algoritmos de *Clustering*. O anexo B é constituído pelas empresas estudadas nesta dissertação e respetivos setores.

Anexo A - Resultados dos índices

Este anexo é constituído por todos os resultados obtidos através do software R de todos os índices aplicados nesta dissertação aos algoritmos *K-Means, PAM, Single Linkage, Average Linkage, Complete Linkage, Diana, Funny, C-Means* tanto com a distância euclidiana como com a distância de *Manhattan.*

K-Means

As tabelas seguintes são constituídas pelos resultados dos índices do algoritmo *K-Means* com a distância euclidiana e com a distância de *Manhattan.*

Tabela 4 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo k-means com a distância euclidiana

	Calinski Harabasz	***Davies Bouldin***	***C index***	***Dunn***	***Gamma***
2	123,61020	0,41276	0,01482	0,32206	-0,97902
3	141,368	0,408	0,027	0,157	-0,923
4	100,9819	0,5606808	0,0525477	0,04124999	-0,7912067
5	176,4781	0,4705019	0,02588214	0,07641999	-0,8739411
6	141,4324	0,5905834	0,03895385	0,06841351	-0,7881075
7	115,1524	0,7511849	0,05046287	0,04901845	-0,7087894
8	194,7102	0,7787328	0,01688903	0,09172388	-0,8939525
9	148,2353	0,2724074	0,03463425	0,06572039	-0,7723487
10	146,0993	0,3525383	0,01862697	0,09967144	-0,8821231
11	129,6517	0,597287	0,02884096	0,01610273	-0,8056966
12	114,2459	0,3653381	0,0303814	0,01610273	-0,7943106
13	100,8265	0,2540005	0,03373125	0,02944021	-0,7703302

Tabela 5 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo k-means com a distância euclidiana

	G plus	**GDI12**	**McClain Rao**	**PBM**
2	0,37393	1,17035	0,20846	85,64272
3	0,462	0,571	0,175	124,486
4	0,3366484	0,1499011	0,205565	102,7154
5	0,3299008	0,2123302	0,1495542	176,4244
6	0,2534843	0,1900845	0,1730734	139,0218
7	0,2087918	0,136196	0,200847	113,5149
8	0,213732	0,1900845	0,1121595	137,0788
9	0,1750779	0,136196	0,1576812	113,2497
10	0,179629	0,2065546	0,1126145	100,9091
11	0,1325698	0,03337057	0,1376295	93,6703
12	0,1252548	0,03337057	0,1416643	81,75494
13	0,1149571	0,06101057	0,1511267	70,57901

Tabela 6 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo k-means com a distância euclidiana

	Point Biserial	**Ray Turi**	**SD**	**Xie Beni**
2	-3,53107	0,03872	13,40176258	0,76919
3	-2,559	0,181	8,724209581	1,581
4	-1,826519	1,421989	8,403338693	20,97619
5	-1,879876	0,6293585	5,44498115	9,283855
6	-1,540182	1,618951	6,810525795	11,22917
7	-1,339037	7,15663	12,0046622	21,69592
8	-1,431686	0,805383	6,346837039	5,586201
9	-1,240713	3,978289	11,78893119	12,06051
10	-1,286507	3,152947	11,79639356	4,580127
11	-1,065967	56,59121	47,70297295	171,6872
12	-1,030153	56,22914	50,51812902	170,5887
13	-0,9768168	33,7992	41,86834187	50,97095

Tabela 7 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo k-means com a distância Manhattan

	Calinski Harabasz	**Davies Bouldin**	**C index**	**Dunn**	**Gamma**
2	116,9604	0,4832356	0,0279815	0,212552	-0,9527898
3	141,3676	0,5110863	0,02690282	0,1569914	-0,9232523
4	191,0571	0,3710294	0,008843011	0,2498763	-0,9699468
5	168,0059	0,5024949	0,02409896	0,04418574	-0,879187
6	58,24274	0,5683109	0,08224213	0,02645919	-0,6161807
7	108,6311	0,8424548	0,03911483	0,01885444	-0,7786736
8	91,62603	0,918259	0,04952876	0,04211392	-0,7060755
9	156,0886	0,4455228	0,0239824	0,09983917	-0,8512944
10	72,63639	0,4107101	0,04623909	0,01368386	-0,7270142
11	117,8921	0,3746674	0,03556839	0,01557675	-0,7641652
12	124,891	0,384054	0,02339937	0,06572039	-0,8435543
13	108,6921	0,4408407	0,02995491	0,02135419	-0,7996162

Tabela 8 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo k-means com a distância Manhattan

	G plus	GDI12	McClain Rao	PBM
2	0,4396988	0,8892284	0,2065624	73,61453
3	0,4615749	0,5705017	0,1746817	124,4855
4	0,4590339	0,8080976	0,1447627	165,0629
5	0,3173903	0,1428962	0,1439593	169,2128
6	0,2291117	0,09615181	0,278983	55,08053
7	0,2333548	0,06097509	0,1725187	98,34974
8	0,1852703	0,136196	0,198676	86,17344
9	0,2243134	0,2069022	0,1326447	110,6777
10	0,1550619	0,04425347	0,1903402	66,36827
11	0,1703444	0,03228057	0,1601475	85,76749
12	0,1264584	0,136196	0,1203855	94,36596
13	0,1102074	0,04425347	0,1388455	78,96598

Tabela 9 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo k-means com a distância Manhattan

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,350335	0,05115729	18,9663483	0,9723708
3	-2,558949	0,1813405	9,85952430	1,580906
4	-2,560875	0,1292595	6,05340837	0,8036574
5	-1,831428	0,6892793	6,52807849	21,48323
6	-1,315236	8,313575	12,4740616	50,02215
7	-1,46081	11,51254	15,4976215	114,4466
8	-1,245403	7,447048	13,8243538	22,57634
9	-1,461188	1,325306	6,95560425	4,968974
10	-1,121126	96,68861	49,2161951	196,5547
11	-1,219487	59,28518	40,6365256	201,3693
12	-1,046883	3,095564	13,2889006	9,384458
13	-0,9594597	44,27345	47,8133271	90,00184

PAM

As tabelas seguintes são constituídas pelos resultados dos índices do algoritmo *PAM* com as distâncias euclidiana e *Manhattan.*

Tabela 10 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo PAM com a distância euclidiana

	Calinski Harabasz	Davies Bouldin	C index	Dunn	Gamma
2	116,9604	0,4832356	0,0279815	0,212552	-0,9527898
3	141,3676	0,469143	0,02690282	0,1569914	-0,9232523
4	194,4483	0,3645805	0,01362107	0,2728122	-0,9538233
5	171,777	0,3433304	0,03315859	0,07168225	-0,845719
6	236,3406	0,3400205	0,0136907	0,1024584	-0,9240181
7	251,3181	0,3653145	0,01143461	0,1122088	-0,9332942
8	279,218	0,1528246	0,01042301	0,1502712	-0,9372863
9	290,83	0,1544543	0,007217616	0,1746156	-0,9530978
10	289,5775	0,1591902	0,006762292	0,1746156	-0,9556813
11	332,044	0,1568082	0,004446214	0,1746156	-0,9687176
12	344,3207	0,1605586	0,00807354	0,1701518	-0,9419743
13	407,2483	0,1141902	0,00656037	0,2540948	-0,9514509

Tabela 11 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo PAM com a distância euclidiana

	G plus	GDI12	McClain Rao	PBM
2	0,4396988	0,8892284	0,2065624	73,61453
3	0,4615749	0,5705017	0,1746817	124,4855
4	0,4634959	0,8080976	0,1547855	170,3657
5	0,3347761	0,2123302	0,1657366	174,0929
6	0,3218198	0,2123302	0,1214444	175,9931
7	0,3201825	0,2244176	0,1160964	193,0557
8	0,3192342	0,3880403	0,1136523	198,6777
9	0,2912791	0,3880403	0,100804	190,9706
10	0,2898891	0,4444331	0,09943968	191,6791
11	0,2772367	0,4444331	0,09135435	202,8761
12	0,202498	0,4330718	0,08623394	220,3531
13	0,1992235	0,5081895	0,08139551	236,5954

Tabela 12 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo PAM com a distância euclidiana

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,350335	0,05115729	2,282150565	0,9723708
3	-2,558949	0,1813405	1,735896912	1,580906
4	-2,594267	0,09066574	1,178723954	0,7904135
5	-1,882624	0,6457931	2,640406857	9,526287
6	-1,875858	0,3715478	2,775986491	5,480813
7	-1,874766	0,2838743	2,532328152	4,187514
8	-1,873506	0,2130304	2,305532569	3,142475
9	-1,764163	0,2309062	2,74358569	2,559154
10	-1,759281	0,1993683	2,849169655	2,209618
11	-1,711995	0,1826446	3,216270036	1,676712
12	-1,40522	0,3981039	5,162224329	1,492612
13	-1,394594	0,2971896	5,144727597	1,114253

Tabela 13 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo PAM com a distância Manhattan

	Calinski Harabasz	Davies Bouldin	C index	Dunn	Gamma
2	116,9604	0,4832356	0,0279815	0,212552	-0,9527898
3	141,3676	0,469143	0,02690282	0,1569914	-0,9232523
4	194,4483	0,3645805	0,01362107	0,2728122	-0,9538233
5	171,5507	0,3433304	0,03504511	0,04824149	-0,8339923
6	235,8107	0,3400205	0,01507937	0,06895355	-0,9143493
7	250,5857	0,3653145	0,01277432	0,07551549	-0,9239011
8	278,1403	0,1528246	0,01174525	0,1011311	-0,9280209
9	289,4567	0,1544543	0,008288447	0,1175147	-0,9450944
10	266,7349	0,1591902	0,01041766	0,1175147	-0,9287001
11	271,3572	0,1209713	0,009438336	0,123206	-0,934442
12	321,9376	0,1043976	0,006229098	0,1305738	-0,954558
13	369,4798	0,1095978	0,005534473	0,1708729	-0,9588989

Tabela 14 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo PAM com a distância de Manhattan

	G plus	GDI12	McClain Rao	PBM
2	0,4396988	0,8892284	0,2065624	73,61453
3	0,4615749	0,5705017	0,1746817	124,4855
4	0,4634959	0,8080976	0,1547855	170,3657
5	0,3271044	0,1428962	0,1689536	175,9018
6	0,3138564	0,1428962	0,1233241	177,979
7	0,3121867	0,151031	0,1177983	195,2425
8	0,3112222	0,2611475	0,1152717	201,1127
9	0,2829753	0,2611475	0,1019587	193,444
10	0,2473161	0,2611475	0,1008652	173,8041
11	0,2441875	0,2611475	0,09769981	169,4455
12	0,2307895	0,2611475	0,08652825	183,862
13	0,2292698	0,3417457	0,08434471	208,2423

Tabela 15 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo PAM com a distância Manhattan

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,350335	0,05115729	2,402325	0,9723708
3	-2,558949	0,1813405	1,79441	1,580906
4	-2,594267	0,09066574	1,20377	0,7904135
5	-1,847017	0,67569	2,707363	21,05968
6	-1,841016	0,3891093	2,84375	12,12763
7	-1,840016	0,2974922	2,595189	9,272139
8	-1,838795	0,2234618	2,360232	6,964784
9	-1,72917	0,2319881	2,749103	5,676864
10	-1,582459	0,7229396	5,011509	5,291622
11	-1,57181	0,6177673	5,113922	4,521804
12	-1,523647	0,4570192	5,272971	3,345194
13	-1,518822	0,3515835	5,451261	2,573448

Método *Single Linkage*

As tabelas seguintes são constituídas pelos resultados dos índices do algoritmo *Single Linkage* com as distâncias euclidiana e *Manhattan.*

Tabela 16 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo Single Linkage com a distância euclidiana

	Calinski Harabasz	Davies Bouldin	C index	Dunn	Gamma
2	50,18797	0,3193838	0,04381592	0,2998282	-0,9526358
3	25,03041	0,450135	0,05357652	0,2684039	-0,9380606
4	16,55085	0,2053017	0,05851204	0,2495347	-0,9300342
5	43,56485	0,1883417	0,02259009	0,3403581	-0,9658336
6	60,13647	0,1859771	0,01344422	0,4726854	-0,9613101
7	49,50335	0,1882393	0,01451064	0,3434598	-0,9578186
8	41,90742	0,1335823	0,01671495	0,2624574	-0,9503831
9	35,84269	0,1067969	0,01793898	0,2555527	-0,9460113
10	152,2541	0,07117713	0,002968745	0,5377361	-0,9878849
11	159,4903	0,05797991	0,004768643	0,4327557	-0,9791529
12	149,7142	0,05139236	0,006059152	0,3391038	-0,9717816
13	134,447	0,05139236	0,006475699	0,2988413	-0,9692794

Tabela 17 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo Single Linkage com a distância euclidiana

	G plus	GDI12	McClain Rao	PBM
2	0,2484387	2,082777	0,2501076	73,89124
3	0,2504407	1,864487	0,2537841	43,04337
4	0,251312	1,73341	0,2556875	25,86306
5	0,4054946	2,192276	0,2096606	41,27055
6	0,4605983	2,435152	0,1941206	46,31186
7	0,4604686	1,769415	0,1947795	36,63737
8	0,4600714	1,352112	0,1961622	29,7239
9	0,4597026	1,316541	0,196956	24,90211
10	0,4528456	2,146104	0,1324421	95,71742
11	0,4382236	2,108969	0,1304495	96,83611
12	0,4307182	1,773091	0,1307056	91,25599
13	0,429166	1,562568	0,1310775	84,28934

Tabela 18 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo Sinle Linkage com a distância euclidiana

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,138323	0,05215849	1,635364	0,757241
3	-3,102281	1,160305	1,907682	0,930859
4	-3,083752	1,063803	1,580454	1,063803
5	-3,473801	0,416736	1,28996	0,542986
6	-3,340085	0,251757	1,206642	0,397401
7	-3,329423	0,739541	1,891549	0,73954
8	-3,3077	0,9418202	2,270254	1,24332
9	-3,296073	1,29825	2,831071	1,29825
10	-2,526101	0,2830438	2,733374	0,287386
11	-2,437998	0,2353048	2,635939	0,368888
12	-2,39562	0,317312	3,186758	0,560907
13	-2,387303	0,7090714	4,792928	0,7090714

Tabela 19 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo Single Linkage com a distância Manhattan

	Calinski Harabasz	Davies Bouldin	C index	Dunn	Gamma
2	50,18797	0,3193838	0,04381592	0,2998282	-0,9526358
3	85,15773	0,3968447	0,01805886	0,3403581	-0,9749063
4	57,66804	0,4360337	0,02103316	0,3403581	-0,9692217
5	43,56485	0,1883417	0,02259009	0,3403581	-0,9658336
6	60,13647	0,1859771	0,01344422	0,4726854	-0,9613101
7	49,50335	0,1882393	0,01451064	0,3434598	-0,9578186
8	184,6486	0,1691541	0,002434128	0,5377361	-0,9902871
9	168,9417	0,1383925	0,002713337	0,5377361	-0,988973
10	152,2541	0,07117713	0,002968745	0,5377361	-0,9878849
11	159,4903	0,05797991	0,004768643	0,4327557	-0,9791529
12	149,7142	0,05139236	0,006059152	0,3391038	-0,9717816
13	134,447	0,05139236	0,006475699	0,2988413	-0,9692794

Tabela 20 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo Single Linkage com a distância Manhattan

	G plus	GDI12	McClain Rao	PBM
2	0,2484387	2,082777	0,2501076	73,89124
3	0,4037884	1,657757	0,2078283	78,14794
4	0,4050123	1,752129	0,2090082	58,44087
5	0,4054946	2,192276	0,2096606	41,27055
6	0,4605982	2,435152	0,1941206	46,31186
7	0,4604686	1,769415	0,1947795	36,63737
8	0,4559094	1,691447	0,1326916	118,4871
9	0,4539398	1,932618	0,1324153	104,6114
10	0,4528456	2,146104	0,1324421	95,71742
11	0,4382236	2,108969	0,1304495	96,83611
12	0,4307182	1,773091	0,1307056	91,25599
13	0,429166	1,562568	0,1310775	84,28934

Tabela 21 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo Sinle Linkage com a distância Manhattan

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,138322	0,05215849	1,635364	0,7572409
3	-3,503894	0,1221686	1,105055	0,5813473
4	-3,484216	0,4631629	1,480637	0,5601305
5	-3,473801	0,416736	1,28996	0,5429864
6	-3,340085	0,251757	1,206642	0,3974008
7	-3,329423	0,739541	1,891549	0,739541
8	-2,545583	0,1775062	1,841235	0,3255484
9	-2,53306	0,2197239	2,245968	0,301477
10	-2,526101	0,2830438	2,733374	0,2873856
11	-2,437998	0,2353048	2,635939	0,3688884
12	-2,39562	0,317312	3,186758	0,560907
13	-2,387303	0,7090714	4,792928	0,7090714

Método *Complete Linkage*

As tabelas seguintes são constituídas pelos resultados dos índices do algoritmo *Complete Linkage* com as distâncias euclidiana e *Manhattan*.

Tabela 22 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo Complete Linkage com a distância euclidiana

	Calinski Harabasz	Davies Bouldin	C index	Dunn	Gamma
2	123,6102	0,4127637	0,01482321	0,3220571	-0,979024
3	85,15773	0,3968447	0,01805886	0,3403581	-0,9749063
4	194,4483	0,3645805	0,01362107	0,2728122	-0,9538233
5	223,4024	0,3661826	0,003843138	0,3899416	-0,985841
6	211,3444	0,3937631	0,002776114	0,4270503	-0,9893523
7	206,8855	0,1644885	0,002343865	0,5377361	-0,9906049
8	262,6048	0,1528246	0,01584441	0,1464297	-0,9019357
9	282,2456	0,1538687	0,01209599	0,1701518	-0,9209782
10	280,9483	0,1135801	0,0113579	0,1890603	-0,9249539
11	288,5348	0,1157463	0,01090856	0,2092157	-0,9274464
12	316,4601	0,1114513	0,009466954	0,2095081	-0,9349632
13	382,0117	0,1132154	0,0060563	0,2540948	-0,9562021

Tabela 23 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo Compete Linkage com a distância euclidiana

	G plus	GDI12	McClain Rao	PBM
2	0,393442	1,170345	0,2084606	85,64272
3	0,4037884	1,657757	0,2078283	78,14794
4	0,4634959	0,8080976	0,1547855	170,3657
5	0,4581261	0,8080976	0,1360033	147,3075
6	0,4573318	0,8541006	0,133845	146,3481
7	0,456805	1,476825	0,1328977	134,3717
8	0,3070115	0,3781206	0,1241499	192,9793
9	0,2864889	0,3781206	0,111855	190,4031
10	0,2835791	0,3781206	0,1095856	180,6839
11	0,282181	0,5081895	0,1082665	187,9663
12	0,2614031	0,5081895	0,1010675	195,9361
13	0,2103318	0,5081895	0,08199941	215,7881

Tabela 24 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo Complete Linkage com a distância euclidiana

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3.531069	0,03872417	1,392852	0,7691854
3	-3,503894	0,1221686	0,961458	0,5813473
4	-2,594267	0,09066574	1,058981	0,7904135
5	-2,558689	0,08220759	1,239337	0,5111171
6	-2,554338	0,08288327	1,234414	0,4218288
7	-2,551547	0,06377185	1,074143	0,3496801
8	-1,816377	0,2616136	2,440252	3,515528
9	-1,738299	0,2065897	2,494311	2,776124
10	-1,727707	0,1784335	2,563904	2,397766
11	-1,722764	0,1510395	2,657294	2,029649
12	-1,640835	0,1800769	3,336928	1,623045
13	-1,440671	0,1614176	3,7071	1,187464

Tabela 25 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo Complete Linkage com a distância Manhattan

	CalinskiHarabasz	Davies Bouldin	C index	Dunn	Gamma
2	116,9604	0,4832356	0,0279815	0,212552	-0,9527898
3	119,4111	0,4288884	0,01090995	0,4657174	-0,9695904
4	180,9862	0,3755125	0,01784635	0,08197656	-0,9264897
5	205,7797	0,3697691	0,01234807	0,1279276	-0,9445896
6	191,5916	0,397079	0,01108431	0,1372406	-0,9487857
7	205,3791	0,3606363	0,008134914	0,1401018	-0,9558226
8	215,357	0,1485591	0,007484629	0,2148398	-0,9580641
9	262,2392	0,1544762	0,01334055	0,1701518	-0,9102308
10	257,0691	0,1166302	0,01247129	0,1890603	-0,9151004
11	282,6771	0,1098073	0,01078853	0,1890603	-0,925508
12	314,0994	0,1043976	0,008061463	0,1890603	-0,9428082
13	357,8902	0,1095978	0,00736351	0,2474101	-0,9471884

Tabela 26 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo Complete Linkage com a distância Manhattan

	G plus	GDI12	McClain Rao	PBM
2	0,4396988	0,8892284	0,2065624	73,61453
3	0,4559337	1,506125	0,1933568	99,70956
4	0,4329025	0,2651114	0,1523416	161,4907
5	0,4315084	0,2651114	0,1423628	143,1294
6	0,430552	0,2802036	0,1398842	141,4195
7	0,4022565	0,2802036	0,1272236	131,3809
8	0,4015351	0,4844997	0,1257744	128,6357
9	0,2562117	0,3781206	0,1100362	174,9978
10	0,2531317	0,3781206	0,1072859	164,2539
11	0,2469068	0,3781206	0,1017758	175,8904
12	0,225363	0,3781206	0,09039969	183,5926
13	0,2238352	0,4948202	0,08820106	207,3812

Tabela 27 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo Complete Linkage com a distância de Manhattan

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,350335	0,05115729	1,698482	0,9723708
3	-3,393074	0,05865	0,990259	0,5280899
4	-2,398303	0,1568255	1,392347	7,857933
5	-2,398094	0,1025799	1,308413	5,13989
6	-2,394333	0,09115216	1,232969	4,310296
7	-2,24825	0,2111426	2,011708	3,272185
8	-2,246197	0,1678876	1,830258	2,601839
9	-1,613524	0,2891784	2,817805	2,985029
10	-1,602781	0,2536057	2,920727	2,617832
11	-1,580606	0,2690829	3,414475	2,07131
12	-1,498438	0,2124218	3,565358	1,635152
13	-1,49357	0,1645999	3,687039	1,267035

Método *Average Linkage*

As tabelas seguintes são constituídas pelos resultados dos índices do algoritmo *Average Linkage* com as distâncias euclidiana e *Manhattan*.

Tabela 28 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo Average Linkage com a distância euclidiana

	Calinski Harabasz	**Davies Bouldin**	**C index**	**Dunn**	**Gamma**
2	116,9604	0,4832356	0,0279815	0,212552	-0,9527898
3	119,4111	0,4288884	0,01090995	0,4657174	-0,9695904
4	94,15535	0,4083652	0,01195554	0,4726854	-0,9662107
5	223,4024	0,3661826	0,003843138	0,3899416	-0,985841
6	211,3444	0,3937631	0,002776114	0,4270503	-0,9893523
7	206,8855	0,1644885	0,002343865	0,5377361	-0,9906049
8	184,6486	0,1691541	0,002434128	0,5377361	-0,9902871
9	169,3616	0,1226906	0,002698735	0,5377361	-0,9891505
10	173,8779	0,121523	0,004572266	0,4327557	-0,980221
11	321,4344	0,1157463	0,005380801	0,2147043	-0,9639953
12	317,09	0,05797991	0,005192922	0,2147043	-0,9650815
13	373,8514	0,04011314	0,003480914	0,2784518	-0,9745051

Tabela 29 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo Average Linkage com a distância euclidiana

	G plus	**GDI12**	**McClain Rao**	**PBM**
2	0,4396988	0,8892284	0,2065624	73,61453
3	0,4559337	1,506125	0,1933568	99,70956
4	0,4596215	1,506125	0,193413	68,48551
5	0,4581261	0,8080976	0,1360033	147,3075
6	0,4573318	0,8541006	0,133845	146,3481
7	0,456805	1,476825	0,1328977	134,3717
8	0,4559094	1,691447	0,1326916	118,4871
9	0,4539803	1,984835	0,1323943	104,9376
10	0,4394151	1,597342	0,1304517	102,7864
11	0,2946793	0,5215215	0,09702339	197,1171
12	0,2930663	0,6885642	0,09623304	201,7548
13	0,2724344	0,8323924	0,08795396	220,3499

Tabela 30 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo Average Linkage com a distância de euclidiana

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,350335	0,05115729	1,593114	0,9723708
3	-3,393074	0,05865	0,946725	0,5280899
4	-3,361589	0,1664436	1,114313	0,4438746
5	-2,558689	0,08220759	1,219962	0,5111171
6	-2,554338	0,08288327	1,219012	0,4218288
7	-2,551547	0,06377185	1,066967	0,3496801
8	-2,545583	0,1775062	1,823436	0,3255484
9	-2,533148	0,2127247	2,182563	0,3007453
10	-2,444884	0,2130542	2,325182	0,3895161
11	-1,781141	0,1173859	2,467749	1,731594
12	-1,774897	0,1045839	2,562023	1,538085
13	-1,694314	0,1346091	3,413334	1,152003

Tabela 31 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo Average Linkage com a distância Manhattan

	Calinski Harabasz	Davies Bouldin	C index	Dunn	Gamma
2	116,9604	0,4832356	0,0279815	0,212552	-0,9527898
3	119,4111	0,4288884	0,01090995	0,4657174	-0,9695904
4	94,15535	0,4083652	0,01195554	0,4726854	-0,9662107
5	223,4024	0,3661826	0,003843138	0,3899416	-0,985841
6	211,3444	0,3937631	0,002776114	0,4270503	-0,9893523
7	206,8855	0,1644885	0,002343865	0,5377361	-0,9906049
8	184,6136	0,1341915	0,002629566	0,5377361	-0,9892591
9	168,9417	0,1383925	0,002713337	0,5377361	-0,988973
10	173,366	0,1263921	0,004587132	0,4327557	-0,9800383
11	319,4626	0,1201728	0,005401274	0,2147043	-0,9637253
12	367,6129	0,1156647	0,003545554	0,2539008	-0,9737756
13	383,2377	0,03913723	0,003258722	0,3132368	-0,9756506

Tabela 32 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo Average Linkage com a distância Manhattan

	G plus	GDI12	McClain Rao	PBM
2	0,4396988	0,8892284	0,2065624	73,61453
3	0,4559337	1,506125	0,1933568	99,70956
4	0,4596215	1,506125	0,193413	68,48551
5	0,4581261	0,8080976	0,1360033	147,3075
6	0,4573318	0,8541006	0,133845	146,3481
7	0,456805	1,476825	0,1328977	134,3717
8	0,4548435	1,476825	0,1326309	114,0678
9	0,4539398	1,932618	0,1324153	104,6114
10	0,4393746	1,55532	0,1304744	102,4352
11	0,2946388	0,5078015	0,0970694	196,1204
12	0,2742135	0,5078015	0,08848871	207,9076
13	0,2725924	0,8922435	0,08743155	222,6866

Tabela 33 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo Average Linkage com a distância de Manhattan

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,350335	0,05115729	1,55027	0,9723708
3	-3,393074	0,05865	0,929023	0,5280899
4	-3,361589	0,1664436	1,103715	0,4438746
5	-2,558689	0,08220759	1,212712	0,5111171
6	-2,554338	0,08288327	1,213249	0,4218288
7	-2,551547	0,06377185	1,064282	0,3496801
8	-2,538998	0,2373117	2,182557	0,3256087
9	-2,53306	0,2197239	2,230287	0,301477
10	-2,444795	0,2136721	2,347331	0,3906458
11	-1,781032	0,1181043	2,493774	1,742192
12	-1,701575	0,1446646	3,26265	1,328048
13	-1,695457	0,1224312	3,306504	1,123941

Diana

As tabelas seguintes são constituídas pelos resultados dos índices do algoritmo *Diana* com as distâncias euclidiana e *Manhattan*.

Tabela 34 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo Diana com a distância euclidiana

	Calinski Harabasz	Davies Bouldin	C index	Dunn	Gamma
2	123,6102	0,4127637	0,01482321	0,3220571	-0,979024
3	85,15773	0,3968447	0,01805886	0,3403581	-0,9749063
4	194,4483	0,3645805	0,01362107	0,2728122	-0,9538233
5	223,4024	0,3661826	0,003843138	0,3899416	-0,985841
6	211,3444	0,3937631	0,002776114	0,4270503	-0,9893523
7	206,8855	0,1644885	0,002343865	0,5377361	-0,9906049
8	279,218	0,1528246	0,01042301	0,1502712	-0,9372863
9	305,1597	0,1538687	0,00657291	0,1746156	-0,9573317
10	307,6478	0,1135801	0,005824935	0,1940201	-0,9614445
11	321,4344	0,1157463	0,005380801	0,2147043	-0,9639953
12	370,5891	0,1122553	0,003523626	0,2607608	-0,9740673
13	383,2377	0,03913723	0,003258722	0,3132368	-0,9756506

Tabela 35 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo Diana com a distância euclidiana

	G plus	GDI12	McClain Rao	PBM
2	0,393442	1,170345	0,2084606	85,64272
3	0,4037884	1,657757	0,2078283	78,14794
4	0,4634959	0,8080976	0,1547855	170,3657
5	0,4581261	0,8080976	0,1360033	147,3075
6	0,4573318	0,8541006	0,133845	146,3481
7	0,456805	1,476825	0,1328977	134,3717
8	0,3192342	0,3880403	0,1136523	198,6777
9	0,2989386	0,3880403	0,10073	197,2143
10	0,2960612	0,3880403	0,09837778	188,0387
11	0,2946793	0,5215215	0,09702339	197,1171
12	0,274254	0,5215215	0,08843733	209,095
13	0,2725924	0,8922435	0,08743155	222,6866

Tabela 36 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo Diana com a distância euclidiana

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,531069	0,03872417	1,266271	0,7691854
3	-3,503894	0,1221686	0,909958	0,5813473
4	-2,594267	0,09066574	1,02619	0,7904135
5	-2,558689	0,08220759	1,212712	0,5111171
6	-2,554338	0,08288327	1,213249	0,4218288
7	-2,551547	0,06377185	1,064282	0,3496801
8	-1,873506	0,2130304	2,267133	3,142475
9	-1,796441	0,1654357	2,31699	2,440392
10	-1,786007	0,1410812	2,380187	2,081131
11	-1,781141	0,1173859	2,467051	1,731594
12	-1,701688	0,1435102	3,224817	1,31745
13	-1,695457	0,1224312	3,306504	1,123941

Tabela 37 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo Diana com a distância Manhattan

	Calinski Harabasz	Davies Bouldin	C index	Dunn	Gamma
2	123,6102	0,4127637	0,01482321	0,3220571	-0,979024
3	85,15773	0,3968447	0,01805886	0,3403581	-0,9749063
4	194,4483	0,3645805	0,01362107	0,2728122	-0,9538233
5	223,4024	0,3661826	0,003843138	0,3899416	-0,985841
6	211,3444	0,3937631	0,002776114	0,4270503	-0,9893523
7	206,8855	0,1644885	0,002343865	0,5377361	-0,9906049
8	279,218	0,1528246	0,01042301	0,1502712	-0,9372863
9	307,373	0,1502369	0,006311912	0,1746156	-0,9582337
10	308,662	0,1103103	0,005530314	0,1940201	-0,9623944
11	298,9567	0,1152227	0,004806095	0,1940201	-0,9661753
12	322,7978	0,1199542	0,004296484	0,2539008	-0,9691538
13	335,467	0,1239557	0,003965206	0,2539008	-0,9708099

Tabela 38 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo Diana com a distância Manhattan

	G plus	GDI12	McClain Rao	PBM
2	0,393442	1,170345	0,2084606	85,64272
3	0,4037884	1,657757	0,2078283	78,14794
4	0,4634959	0,8080976	0,1547855	170,3657
5	0,4581261	0,8080976	0,1360033	147,3075
6	0,4573318	0,8541006	0,133845	146,3481
7	0,456805	1,476825	0,1328977	134,3717
8	0,3192342	0,3880403	0,1136523	198,6777
9	0,2920451	0,3880403	0,09875815	195,2991
10	0,2890948	0,3880403	0,09630715	185,2274
11	0,280568	0,3880403	0,09292723	174,9912
12	0,2791496	0,5078015	0,09137689	189,4232
13	0,2681467	0,5078015	0,08832835	193,16

Tabela 39 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo Diana com a distância Manhattan

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,531069	0,03872417	1,438369	0,7691854
3	-3,503894	0,1221686	0,979977	0,5813473
4	-2,594267	0,09066574	1,070773	0,7904135
5	-2,558689	0,08220759	1,248911	0,5111171
6	-2,554338	0,08288327	1,242024	0,4218288
7	-2,551547	0,06377185	1,077689	0,3496801
8	-1,873506	0,2130304	2,278357	3,142475
9	-1,768964	0,1772569	2,427678	2,423024
10	-1,758451	0,1517503	2,514473	2,074361
11	-1,725098	0,1909018	3,073756	1,860622
12	-1,720278	0,155039	3,190314	1,511086
13	-1,676572	0,1776369	3,851053	1,283003

C-Means

As tabelas seguintes são constituídas pelos resultados dos índices do algoritmo C-*Means* com as distâncias euclidiana e *Manhattan*.

Tabela 40 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo C-Means com a distância euclidiana

	Calinski Harabasz	Davies Bouldin	C index	Dunn	Gamma
2	123,6102	0,4127637	0,01482321	0,3220571	-0,979024
3	141,3676	0,5110863	0,02690282	0,1569914	-0,9232523
4	198,9001	0,4655535	0,009983632	0,2908433	-0,9654518
5	171,777	0,470831	0,03315859	0,07168225	-0,845719
6	236,3406	0,4754241	0,0136907	0,1024584	-0,9240181
7	219,8394	0,4025675	0,01131683	0,06895355	-0,9334536
8	193,4815	0,7100989	0,01608959	0,06913288	-0,9001039
9	169,1502	0,4117685	0,01765747	0,09983917	-0,8897641
10	146,2193	0,3374657	0,02586228	0,06572039	-0,8266687
11	143,2382	0,4229829	0,02176407	0,06572039	-0,8563748
12	127,9174	0,3827477	0,02171582	0,06572039	-0,8582038
13	113,1883	0,4500076	0,02483248	0,06572039	-0,8351497

Tabela 41 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo C-Means com a distância euclidiana

	G plus	GDI12	McClain Rao	PBM
2	0,393442	1,170345	0,2084606	85,64272
3	0,4615749	0,5705017	0,1746817	124,4855
4	0,4615709	0,8080976	0,1477676	172,2238
5	0,3347761	0,2123302	0,1657366	174,0929
6	0,3218198	0,2123302	0,1214444	175,9931
7	0,2865943	0,1428962	0,1097871	154,5886
8	0,2223884	0,1432679	0,1113687	134,1483
9	0,1866765	0,2069022	0,110671	120,2099
10	0,1449547	0,136196	0,129973	109,4963
11	0,1273379	0,136196	0,1154529	110,3237
12	0,1206632	0,136196	0,1144526	99,01926
13	0,1078244	0,136196	0,1224381	87,56016

Tabela 42 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo C-Means com a distância euclidiana

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,531069	0,03872417	4,962011	0,7691854
3	-2,558949	0,1813405	3,616284	1,580906
4	-2,58051	0,1041975	2,238003	0,7736761
5	-1,882624	0,6457931	3,269649	9,526287
6	-1,875858	0,3715478	3,308376	5,480813
7	-1,735194	0,3491957	3,529556	10,53914
8	-1,467879	0,6991403	4,954867	9,894684
9	-1,317539	1,225115	6,713927	4,593324
10	-1,126291	3,472163	11,40199	10,52615
11	-1,053165	3,083382	11,89101	9,347528
12	-1,021795	3,023654	12,4978	9,166456
13	-0,956261	3,922869	15,0926	9,131422

Tabela 43 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo C-Means com a distância Manhattan

	Calinski Harabasz	Davies Bouldin	C index	Dunn	Gamma
2	39,013	0,7117674	0,1979148	0,08562896	-0,6671524
3	119,4111	0,4288884	0,01090995	0,4657174	-0,9695904
4	12,81607	0,7215967	0,2996197	0,02014204	-0,2534733
5	9,342499	0,5788543	0,3250443	0,01443182	-0,1397988
6	45,47622	0,5243014	0,0809175	0,02012459	-0,6622803
7	29,23199	0,4828803	0,09801659	0,01922067	-0,6283224
8	66,81159	0,5968759	0,0389582	0,0293416	-0,8070075
9	117,1699	0,5091441	0,02793979	0,06443224	-0,8337139
10	69,39782	0,6065801	0,06493718	0,01201044	-0,610337
11	47,69289	0,5808806	0,03430932	0,04536931	-0,8431724
12	42,25925	0,5346865	0,06073846	0,007884047	-0,6530691
13	112,2723	0,7710106	0,02402719	0,0245561	-0,8419148

Tabela 44 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo C-Means com a distância Manhattan

	G plus	GDI12	McClain Rao	PBM
2	0,4172796	0,4406186	0,3991097	22,69247
3	0,4559337	1,506125	0,1933568	99,70956
4	0,2570465	0,1036444	0,7480077	8,284344
5	0,2190004	0,07426144	0,8528815	5,741561
6	0,2180845	0,0855779	0,2785745	43,90291
7	0,2120258	0,0804112	0,3235079	28,05132
8	0,2299101	0,09691407	0,1716826	55,6891
9	0,2049945	0,136196	0,1409318	88,68583
10	0,1643992	0,03337057	0,2421794	57,08804
11	0,2040948	0,1498529	0,1577253	46,50357
12	0,1213643	0,02462029	0,2337178	36,60479
13	0,1196054	0,05088897	0,121492	78,51674

Tabela 45 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo C-Means com a distância Manhattan

	Point Biserial	Ray Turi	SD	Xie Beni
2	-1,813562	0,2281362	8,422279	6,887621
3	-3,393074	0,05865	2,899417	0,5280899
4	-0,5226612	17,5499	8,378033	121,7275
5	-0,2848276	78,15542	13,8673	236,9346
6	-1,250002	8,608266	8,700423	62,33239
7	-1,157406	12,61237	9,605758	75,88765
8	-1,432496	3,826951	6,78	31,55203
9	-1,371119	2,519921	8,355492	15,16217
10	-1,127144	119,0252	53,35343	361,1001
11	-1,332902	3,128561	6,884464	11,72992
12	-0,939472	140,2569	56,42044	476,3997
13	-1,013142	4,315355	15,55519	65,93024

Funny

As tabelas seguintes são constituídas pelos resultados dos índices do algoritmo *Funny* com as distâncias euclidiana e *Manhattan*.

Tabela 46 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo Funny com a distância euclidiana

	Calinski Harabasz	Davies Bouldin	C index	Dunn	Gamma
2	116,9604	0,4832356	0,0279815	0,212552	-0,9527898
3	141,3676	0,469143	0,02690282	0,1569914	-0,9232523
4	100,9177	0,469143	0,05469313	0,02776086	-0,7783909
5	176,2398	0,3420208	0,02761807	0,05142995	-0,8629119
6	183,4284	0,3312363	0,01990568	0,06760202	-0,8953173
7	182,7584	0,3575894	0,01741284	0,06760202	-0,905888
8	191,614	0,3661349	0,01694328	0,09983917	-0,8946589
9	163,7441	0,7936451	0,0230771	0,06572039	-0,8476009
10	144,5832	0,8007609	0,02719842	0,06572039	-0,818507
11	144,5832	0,8007609	0,02719842	0,06572039	-0,818507
12	112,5626	0,5337245	0,02786124	0,06572039	-0,8144587
13	118,1319	0,204267	0,02466512	0,06572039	-0,835578

Tabela 47 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo Funny com a distância euclidiana

	G plus	GDI12	McClain Rao	PBM
2	0,4396988	0,8892284	0,2065624	73,61453
3	0,4615749	0,5705017	0,1746817	124,4855
4	0,3291226	0,1008821	0,2099047	103,7985
5	0,3221156	0,1428962	0,1523099	178,2626
6	0,2911495	0,1428962	0,1308431	148,5531
7	0,2893825	0,151031	0,1247479	159,7846
8	0,2177967	0,2069022	0,1129436	137,7396
9	0,1763342	0,136196	0,1250304	120,4842
10	0,1421624	0,136196	0,1336501	109,5375
11	0,1421624	0,136196	0,1336501	109,5375
12	0,1353823	0,136196	0,1349278	85,87256
13	0,1325617	0,136196	0,124919	99,59311

Tabela 48 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo Funny com a distância euclidiana

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,350335	0,05115729	4,436282	0,9723708
3	-2,558949	0,1813405	2,784755	1,580906
4	-1,790109	1,486799	4,055531	46,34006
5	-1,844549	0,6585161	3,213704	20,52441
6	-1,732116	0,4964823	3,207174	15,4742
7	-1,731362	0,4048653	2,829475	12,61871
8	-1,447652	1,277437	6,202829	4,789496
9	-1,266653	1,818697	7,704181	10,94296
10	-1,111301	3,510626	11,19422	10,64275
11	-1,111301	3,510626	11,19422	10,64275
12	-1,079654	3,42763	12,68305	10,39115
13	-1,072538	2,888232	10,68089	8,755913

Tabela 49 - Resultados dos índices Calinski Harabasz, Davies Bouldin, C index, Dunn e Gamma do algoritmo Funny com a distância Manhattan.

	Calinski Harabasz	Davies Bouldin	C index	Dunn	Gamma
2	116,9604	0,4832356	0,0279815	0,212552	-0,9527898
3	141,3676	0,469143	0,02690282	0,1569914	-0,9232523
4	100,9177	0,469143	0,05469313	0,02776086	-0,7783909
5	176,2398	0,3420208	0,02761807	0,05142995	-0,8629119
6	192,1961	0,3329206	0,01797454	0,06895355	-0,902836
7	193,5148	0,3590497	0,01546142	0,0694971	-0,9134949
8	187,7999	0,3673982	0,01784803	0,09983917	-0,8890582
9	161,5823	0,3788907	0,02353045	0,06572039	-0,8442803
10	143,1692	0,4664905	0,02721575	0,06572039	-0,817758
11	125,4141	1,224602	0,03064347	0,06572039	-0,7929007
12	112,5626	0,5337245	0,02786124	0,06572039	-0,8144587
13	115,5111	1,285971	0,02775153	0,06572039	-0,8117406

Tabela 50 - Resultados dos índices G Plus, GDI12, McClain Rao e PBM do algoritmo Funny com a distância Manhattan

	G plus	GDI12	McClain Rao	PBM
2	0,4396988	0,8892284	0,2065624	73,61453
3	0,4615749	0,5705017	0,1746817	124,4855
4	0,3291226	0,1008821	0,2099047	103,7985
5	0,3221156	0,1428962	0,1523099	178,2626
6	0,2923044	0,1428962	0,1265046	152,1074
7	0,2905375	0,151031	0,1203445	164,0116
8	0,2210956	0,2069022	0,1158927	135,1915
9	0,1801356	0,136196	0,1266649	117,1193
10	0,1485129	0,136196	0,1343992	107,2694
11	0,1273217	0,136196	0,142624	91,23722
12	0,1353823	0,136196	0,1349278	85,87256
13	0,1220735	0,136196	0,133325	92,52838

Tabela 51 - Resultados dos índices Point Biserial, Ray Turi, SD e Xie Beni do algoritmo Funny com a distância Manhattan

	Point Biserial	Ray Turi	SD	Xie Beni
2	-3,350335	0,05115729	4,898988	0,9723708
3	-2,558949	0,1813405	3,010049	1,580906
4	-1,790109	1,486799	4,222255	46,34006
5	-1,844549	0,6585161	3,30013	20,52441
6	-1,742473	0,4745634	3,277939	14,79103
7	-1,74178	0,3829464	2,897527	11,93555
8	-1,459392	1,302802	6,255189	4,884599
9	-1,282259	1,842495	7,775857	11,08615
10	-1,13957	3,54456	11,26318	10,74563
11	-1,039138	4,306267	13,25495	10,64802
12	-1,079654	3,42763	12,71471	10,39115
13	-1,019912	3,619978	11,87144	8,951051

Anexo B – Setores das empresas cotadas na bolsa de valores de Lisboa

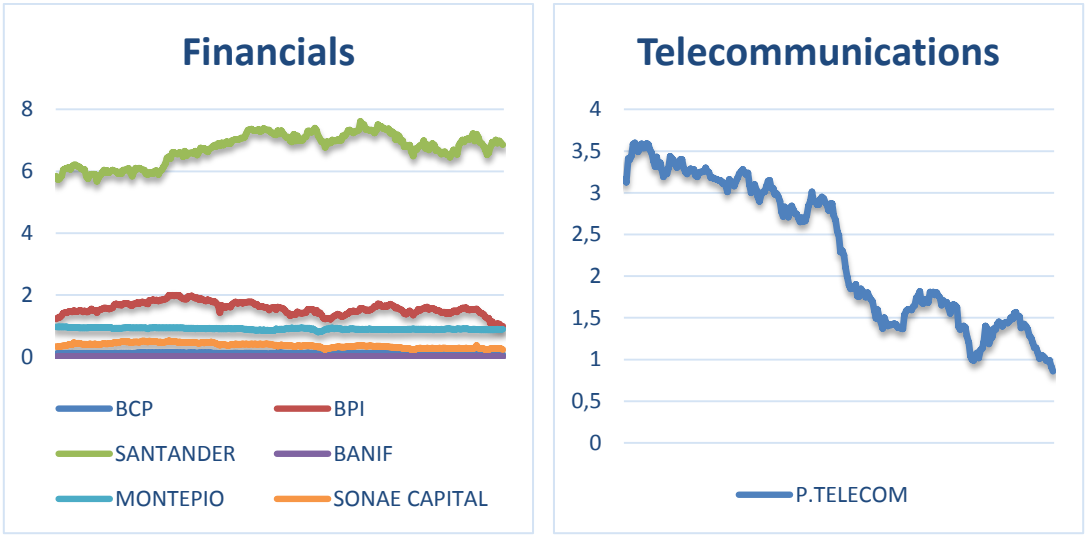


Figura 58 - Empresas dos sectores Financials e Telecommunications

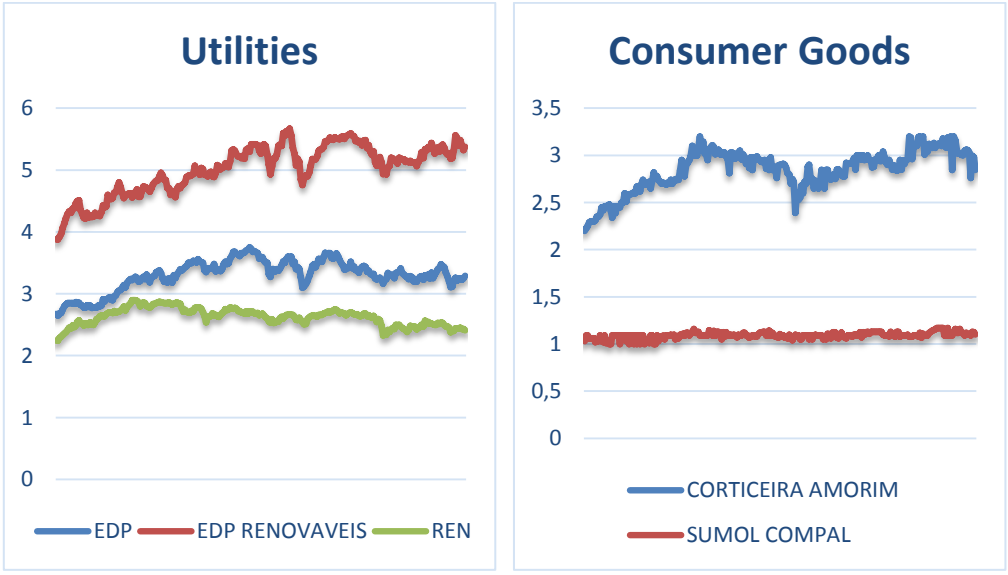


Figura 59 - Empresas dos sectores Utilities e Consumer Goods

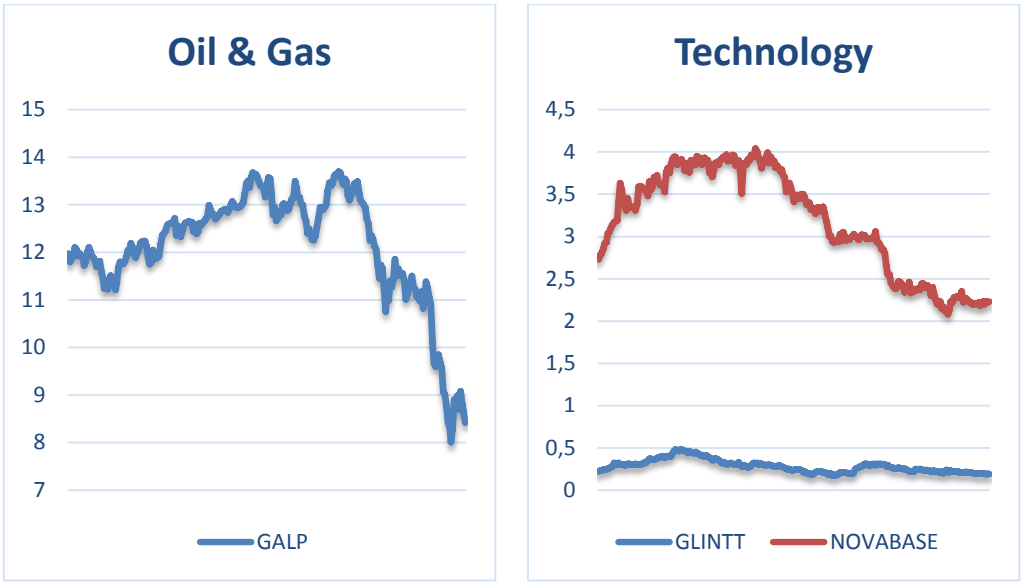


Figura 60 - Empresas dos sectores Oil & Gas e Technology

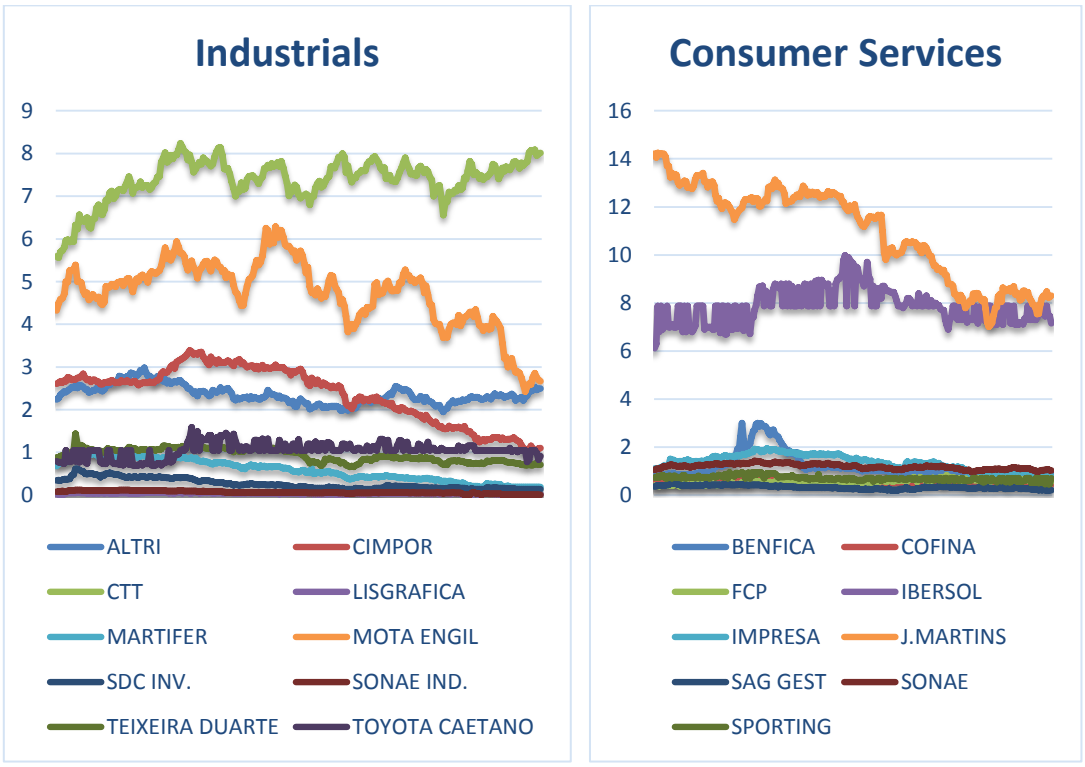


Figura 61 - Empresas dos sectores Industrials e Consumer Services